Intelligent Monitoring and Interpretation of Preterm Physiological Signals using Machine Learning

OKSANA SEMENOVA

INTELLIGENT MONITORING AND INTERPRETATION OF PRETERM PHYSIOLOGICAL SIGNALS USING MACHINE LEARNING

Oksana Semenova

NATIONAL UNIVERSITY OF IRELAND, CORK

SCHOOL OF ENGINEERING

Thesis submitted for the degree of

Doctor of Philosophy

February, 2020

Head of School: Prof. William Marnane

Supervisors: Dr. Gordon Lightbody

Dr. Andriy Temko

Prof. Eugene Dempsey

TABLE OF CONTENTS

Acknowledgements vi
Abstract vii
List of Figures ix
List of Tables xvi
List of Acronyms xviii
Chapter 1: Introduction 1
1.1 Premature birth 1
1.2 Decision support in NICU 4
1.2.1 Computer-based analysis: classical approach 5
1.2.2 Computer-based analysis: end-to-end learning 7
1.3 Aims and scopes of the thesis: Intelligent monitoring of preterm neonates 8
1.4 Thesis layout 11
1.5 List of publications arising from this thesis 12
Chapter 2: Medical background – monitoring of preterm infants with hypotension 14
2.1 Cardiovascular system and blood pressure 14
2.1.1 Neonatal cardiovascular system 14
2.1.2 Blood pressure disorders: hypotension 15
2.1.3 Hypotension and autoregulation 16
2.1.4 Autoregulation in preterm neonates 17
2.1.5 Recording blood pressure 18
2.1.6 Current therapies and outcomes 19
2.2 Measures of physiological and neurodevelopmental health 21
2.3 Heart activity in neonates 24
2.3.1 Recording heart activity 25
2.3.2 ECG artefacts 26
2.4 Brain and electroencephalography 26
2.4.1 Brain function and BP 27
2.4.2 Recording EEG 28
2.4.3 EEG artefacts 29
2.4.4 Characteristics of preterm EEG 30

2.5 Conclusion 33

Chapter 3: Methods - biomedical signal processing and machine learning 34

3.1 Signal processing and feature extraction 34

3.1.1 Blood pressure 34

3.1.2 EEG analysis 35

3.1.3 Heart rate variability (HRV) analysis 37

3.1.4 Time domain HRV features 41

3.1.5 Frequency domain HRV parameters 43

3.1.6 Nonlinear HRV analysis 44

3.1.7 Allan analysis 45

3.2 Modelling interaction between physiological signals 45

3.2.1 Linear measures of interaction: correlation and coherence 45

3.2.2 Nonlinear measure of interaction: mutual information 47

3.2.3 Directionality of interaction: transfer entropy 49

3.3 AUC as a measure of statistical predictive power and classification metric 51

3.4 Machine learning 52

3.4.1 Machine learning paradigms 52

3.4.2 Supervised machine learning: decision trees 54

3.4.3 Supervised machine learning: support vector machine 56

3.4.4 Supervised machine learning: Gaussian mixture model 57

3.4.5 Supervised machine learning: neural networks 59

3.4.6 Convolutional neural network 63

3.5 Bias-variance trade-off in machine learning 68

3.6 Conclusion 70

Chapter 4: Modelling interaction between blood pressure and brain activity 71

4.1 Introduction 71

4.2 Dataset 72

4.3 Exploring the contextual information of signals and features 74

4.4 Modelling of interaction between EEG sub-band power and MAP 77

4.4.1 Linear interaction: correlation and coherence 77

4.4.2 Statistical significance 80

4.4.3 Nonlinear interaction: adjusted mutual information 81

4.4.4 Directionality of interaction 86
4.5 Results of the association of coupling measures with the illness severity score 87
4.6 Decision support tool 90
4.7 Discussion 91
4.7.1 Nonlinear relationship between the MAP and EEG sub-band powers 91
4.7.2 Linear relationship between MAP and EEG sub-band powers 92
4.7.3 Directionality of interaction 92
4.7.4 General discussion 93
4.7.5 Limitations 96
4.8 Conclusions 96
Chapter 5: Prediction of short-term health outcome using multimodal physiological signal analysis and boosted decision trees 98
5.1 Introduction 98
5.2 Dataset 99
5.3 Feature extraction 102
5.4 Exploring the predictive power of HRV characteristics 103
5.5 Exploring the predictive power of EEG characteristics 109
5.6 Machine learning: boosted decision trees 112
5.6.1 Tree models: ensemble methods 112
5.6.2 Gradient boosting with regularization 116
5.6.3 Model selection and performance assessment 121
5.6.4 Statistical inference and out-of-sample predictive modelling 123
5.6.5 Combination of features with boosted decision trees 123
5.7 Feature selection 127
5.8 Decision support tool 130
5.9 Discussion 132
5.10 Conclusion 134
Chapter 6: Automated assessment of 2-years neurodevelopmental outcome using early EEG recordings 135
6.1 Clinical problem and motivation 135
6.2 Dataset 137
6.3 Feature-based approach 138
6.3.1 Feature analysis: exploring feature predictive power 141
6.3.2 Out-of-sample predictive modelling using boosted decision trees 144
6.4 End-to-end deep learning for outcome prediction 145

6.4.1 Related work....145
6.4.2 Methodology....146
6.4.3 Optimization and classification performance....150
6.5 Discussion....151
6.6 Conclusion....154
Chapter 7: Conclusions and future work....155
7.1 Concluding summary and contributions....155
7.1.1 General limitations....159
7.2 Future work....159
References....162

This is to certify that the work I am submitting is my own and has not been submitted for another degree, either at University College Cork or elsewhere. All external references and sources are clearly acknowledged and identified within the contents. I have read and understood the regulations of University College Cork concerning plagiarism.

Oksana Semenova

ACKNOWLEDGEMENTS

I had a privilege of having two supervisors throughout my PhD journey. Firstly, I would like to express my sincere gratitude to Dr Andriy Temko for his continuous support and guidance. At many stages in the course of the research I have greatly benefited from his innovative ideas. I would like to thank Dr Gordon Lightbody for his guidance which helped me in all the time of research and writing of this thesis. I could not have imagined of having better supervisors and mentors for my PhD. Their positive vision and confidence in my research inspired and provided me with an additional self-confidence. I truly enjoyed working in the research environment that stimulated innovative thinking which they have created. I also thank SFI for funding this research.

I would like to thank Prof. Eugene Dempsey and Prof. Geraldine Boylan for taking time out of their busy schedule and helping me to better understand preterm physiology and medical terminology. Their insight and feedback played a vital role in my research. I want also to thank Dr John O'Toole for his support and fruitful collaboration. My sincere thanks go to the staff in the School of Engineering including Ralph O'Flahery, Niam O'Sullivan and others.

I felt very lucky to be surrounded by my amazing colleagues, all the postgraduate students in the Electrical Engineering building and those from Infant. They all have shared this long journey with me and filled it with even more meaning. My gratitude goes to Alison, Brian, Shima, Haiyang, Wentao, James, Rehan, Pablo, George, Yeny, Alex, Robbie, Adrian, Brendan, Sergii, Mark, Maeve, Giorgia, Flora, Jing, Stella, Shiao - thank you all! To my friends and housemates Adhurim, Taejung, Martina, Sarah, Hugo, Cici, Andrea, Sief. Thank you guys for great chats and cooking lessons.

My PhD journey has awarded me with a very special person – my best friend, exceptional man, my fiancé Riccardo Sandon. I am very grateful for his constant support, patience and help. Being far away from my home, Ukraine, Riccardo has become my family here in Ireland. Our journey has only started, and I am looking forward to sharing much more with you. I want to specially thank my parents, Serhii and Liudmula, my brother Bohdan. The truth is that without their daily support, love and encouragement I would not be able to reach this point in my life. I feel blessed to have such a caring family.

Oksana Semenova

ABSTRACT

Every year, more than one in ten babies are born prematurely. In Ireland of the 70000 babies delivered every year, 4500 are born too early. Premature babies are at a higher risk of complications, which may lead to both short-term and long-term adverse health outcomes. The neonatal population is especially vulnerable and a delay in the identification of medical conditions, as well as delays in the initiating the correct treatment, may be fatal. After birth, preterms are admitted to the neonatal intensive care unit (NICU), where a continuous flow of information in the form of physiological signals is available. Physiological signals can assist clinicians in decision making related to the diagnosis and treatment of various diseases. This information, however, can be highly complex, and usually requires expert analysis which may not be available at all times.

The work conducted in this thesis develops a decision support systems for the intelligent monitoring of preterms in the NICU. This will allow for an accurate estimation of the current health status of the preterm neonate as well as the prediction of possible long-term complications. This thesis is comprised of three main work packages (WP), each addressing health complication of preterm on three different stages of life. At the first 12 hours of life the health status is quantified using the clinical risk index for babies (CRIB). This is followed by the assessment of the preterm's well-being at discharge from the NICU using the clinical course score (CCS). Finally, the long-term neurodevelopmental follow-up is assessed using the Bayley III scales of development at two years. This is schematically represented in Figure 1 along with the main findings and contributions.

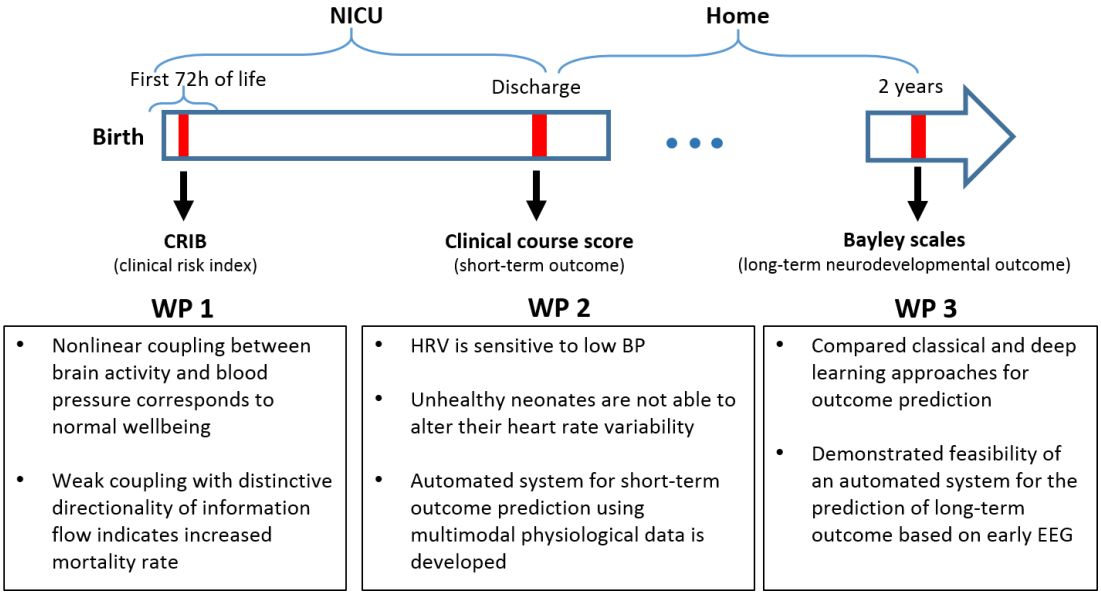


Figure 1. Schematic representation of thesis main findings and contributions.

Low blood pressure (BP) or hypotension is a recognised problem in preterm infants particularly during the first 72 hours of life. Hypotension may cause decreased cerebral perfusion, resulting in deprived oxygen delivery to the brain. Deciding when and whether to treat hypotension relies on our understanding of the relation between BP, oxygenation and brain activity. The electroencephalogram (EEG) is the most commonly used technology to assess the 'brain health' of a newborn. The first WP investigates the relationship between short-term dynamics in BP and EEG energy in the preterm on a large dataset of continuous multi-channel unedited EEG recordings in the context of the health status measured by the CRIB score. The obtained results indicate that a higher risk of mortality for the preterm is associated with a lower level of nonlinear interaction between EEG and BP. The level of coupling between these two systems can potentially serve as an additional source of information when deciding whether or not to intervene in the preterm.

The electrocardiogram (ECG) is also routinely recorded in preterm infants. Analysis of heart rate variability (HRV) provides a non-invasive assessment of both the sympathetic and parasympathetic control of the heart rate. A novel automated objective decision support tool for the prediction of the short-term outcome (CCS) in preterm neonates who may have low BP is proposed in the second WP. Combining multiple HRV features extracted during hypotensive episodes, the classifier achieved an AUC of 0.97 for the task of short-term outcome prediction, using a leave-one-patient-out performance assessment. The developed system is based on the boosted decision tree classifier and allows for the continuous monitoring of the preterm. The proposed system is validated on a large clinically collected dataset of multimodal recordings from preterm neonates.

If the correct treatment is initiated promptly after diagnosis, it can potentially improve the neurodevelopmental outcome of the preterm infant. The third WP presents a pilot study investigating the predictive capability of the early EEG recorded at discharge from the NICU with respect to the 2-year neurodevelopmental outcome using machine learning techniques. Two methods are used: 1) classical feature-based classifier, and 2) end-to-end deep learning. This is a fundamental study in this area, especially in the context of applying end-to-end learning to the preterm EEG for the problem of long-term outcome prediction. It is shown that for the available labelled dataset of 37 preterm neonates, the classical feature-based approach outperformed the end-to-end deep learning technique. A discussion of the obtained result as well as a section highlighting the possible limitations and areas that need to be investigated in the future are provided.

LIST OF FIGURES

Figure 1.1: A preterm neonate being monitored in the NICU at Cork University Maternity Hospital. 8

Figure 1.2: The timeline of an infant's stay in the NICU through to the long-term neurodevelopmental follow-up at 2 years of age. 9

Figure 2.1: Schematic representation of cerebral autoregulation in preterm neonates. The plateau on the curve represents properly functioning autoregulation, where for a range of mean arterial pressure (MAP) the cerebral blood flow (CBF) does not change. This figure is adapted from (Greisen, 2005). 17

Figure 2.2: The infant's course in the NICU through to the neurodevelopmental follow-up at 2 years of age. 21

Figure 2.3: Clinical risk index for babies (CRIB II) (Parry et al., 2003). 23

Figure 2.4: A schematic representation of the ECG waveform and its main components: P wave, QRS complex and T wave. 26

Figure 2.5: The 10-20 international system of the EEG electrodes placement adjusted for the neonates is showed with arrows. 29

Figure 2.6: Preterm neonate, 31 weeks GA. The figure represents a typical discontinuous activity with high amplitude bursts and low amplitude IBI (Pavlidis et al., 2017). 31

Figure 2.7: Preterm neonate, 26 weeks GA. The figure represents synchrony of high amplitude bursts (highlighted). 33

Figure 3.1: Traces of raw systolic (SP) and diastolic (DP) blood pressure along with the corresponding mean arterial pressure (MAP) from the preterm neonate (28 weeks GA). 35

Figure 3.2: Stages of the Pan-Tompkins algorithm. 38

Figure 3.3: An example of the preterm ECG signal going through different stages of the Pan-Tompkins R peak detection. Bandpass cut-off frequencies are 4-30 Hz. 39

Figure 3.4: An example of preterm ECG signal going through different stages of the Pan-Tompkins R peak detection. Bandpass cut-off frequencies are originally proposed 5-15 Hz. 40

Figure 3.5: Schematic representation of the TINN feature computation. *X* is the bin with maximum number of NN intervals, *N* and *M* are determined by finding the interpolated triangle that best fits to the histogram. 42

Figure 3.6: Schematic representation of the signal labelling using histogram binning. 48

Figure 3.7: A contingency table $\mathcal{M}$ with $r \times c$ size, with ai and bj as the row and columns marginal. 49

Figure 3.8: Confusion matrix. 51

Figure 3.9: Description of the ROC curve and the area under the curve (AUC). Each point on the ROC curve represents a pair of sensitivity and specificity values according to the varying threshold. 52

Figure 3.10: A basic structure of a decision tree....55

Figure 3.11: SVM decision boundary for linearly separable class. 56

Figure 3.12: An example of the Gaussian mixture model with two Gaussian components generated for one class of data. 58

Figure 3.13: A diagram of the feedforward neural network with one hidden layer.....59

Figure 3.14: An example of a neuron showing the input and its corresponding weights and the activation function applied to the weighted sum over the inputs from the previous layer. 59

Figure 3.15: Examples of activation functions. 60

Figure 3.16: Visualisation of the possible log loss values given a ground truth of 1. 61

Figure 3.17: The main concept of conventional and residual networks.....62

Figure 3.18: Examples of the filters which perform the vertical and horizontal edge detections.....63

Figure 3.19: Schematic representation of the convolutional operation for a 3D input. The size of the filter is 5x5x3, where each depth of filter is represented by a different matrix of weights. The convolution performed on the 3D input, results in a scalar obtained by multiplication followed by summation.....64

Figure 3.20: An example of the max and average pooling operation with a 2x2 filter and a stride of 2. 65

Figure 3.21: Visual representation of the bias-variance trade-off in machine learning. 69

Figure 3.22: The bias-variance trade-off represented with learning curves. The total error represents the generalisation error of the model. 69

Figure 3.23: Schematic representation of the supervised machine learning workflow.....70

Figure 4.1: The timing of CRIB score along with other illness scores considered in this thesis are assigned to an infant during the course in the neonatal intensive care unit (NICU) through to the neurodevelopmental follow-up at 2 years of age.....72

Figure 4.2: Schematic representation of duration and temporal location of recordings. Each recording is represented with respect to the time of birth (TOB) for each neonate.....73

Figure 4.3: Five minutes of mean arterial pressure (MAP) and 30 seconds of 8-channel raw EEG recording (30 weeks GA).....74

Figure 4.4: An example of a 5-second trace of F4-C4 channel of raw EEG (fs=256 Hz) and its corresponding waveforms in different sub-bands (fs=64 Hz) of preterm with GA of 27 weeks.....75

Figure 4.5: EEG spectrum of a preterm neonate (27 weeks GA) computed over a 10-minute trace of the F4-C4 channel. 75

Figure 4.6: 20 minutes of the sub-band power is represented as a median across 8 EEG bipolar channels for each sub-band (27 weeks GA). 76

Figure 4.7: An example of the raw MAP trace (a) of a preterm neonate (31 weeks GA) and its spectral power (b). The spectrum was computed using the Welch method on the filtered MAP signal (high pass FIR filter with a cut-off frequency of 0.05 Hz) in order to emphasise the Mayer frequency component at 0.1 Hz. The periodogram was computed for each 180-second long segment of the MAP. The obtained result was then averaged to provide a more accurate estimate of the spectral density. To reduce the effect of the spectral leakage a Parzen window (Andriessen et al., 2003) was applied prior to spectral estimation. 76

Figure 4.8: Overview of linear and nonlinear modelling of the interaction between EEG and BP signals. 78

Figure 4.9: Traces of EEG band power and MAP with corresponding coherence between them. Coherence is obtained as a sum of all coherence values within the 30 minutes window (31 weeks GA). 79

Figure 4.10: An example of a 10-hour trace of the EEG band power (0.3-3Hz) and the MAP of preterm (31 weeks GA) with computed cross-correlation over the 30-minute epoch. The reported cross-correlation is a maximum correlation achieved at a certain time lag for a given epoch. The unit of each time lag here corresponds to a 30-second shift. 79

Figure 4.11: Probability density function (PDF) of the Pearson correlation coefficient for surrogates and real data computed on all dataset. Correlation for real data is quantified between MAP and EEG sub-energy (0.3-3 Hz) feature (yellow); correlation between corresponding randomly permuted surrogates of MAP and sub-band energy (0.3-3 Hz) feature (black). 80

Figure 4.12: All correlation values (including insignificant) (red, thin line) and 95th percentile of the correlation calculated using 100 shuffled surrogates (green, bold line) computed for preterm (31 weeks GA)....81

Figure 4.13: One hour of MAP (25 weeks GA) and delta-band energy along with its corresponding labels (dashed line). The shift of a window is 30 sec, which results in 120 values per hour. 81

Figure 4.14: The effect of the number of labels on AMI and MI values. Interaction is calculated between the MAP and the EEG power in the 0.3-3 Hz sub-band for one preterm infant (28 weeks GA). Every value is obtained as a mean across all epochs for a given number of labels. 82

Figure 4.15: A scatter plot of conventional mutual information (MI), adjusted mutual information (AMI) and Pearson correlation. This plot represents measures of interaction between MAP and EEG (0.3-3 Hz) sub-band energy computed for each 30 min window from one preterm (30 weeks GA). 83

Figure 4.16: Trace of artificially generated toy data. $y0 = 2.8 * sin2\pi * 1000 + 1.2 * sin2\pi * 3000x + 0.2 * sin(2\pi * 500x)$ in blue and $y = y0 + N(0,1)$ in green. Pearson correlation coefficient between y_0 and y is equal to 0.91....84

Figure 4.17: The probability density function of adjusted mutual information obtained from all data (25 subjects). AMI for real data (MAP and sub-band energy) (green) and its permuted surrogates (black); toy data (artificially created

correlated signals) (red) and its permuted surrogates (grey). The distributions of the surrogates are overlapped and clipped. 84

Figure 4.18: Artificially generated data with different levels of noise added. 85

Figure 4.19: Effect of noise on AMI (a) and correlation (b). Baselines of zero coupling (blue, zero centred) for both measures are represented as Pearson correlation and AMI between two Gaussian independent and identically distributed processes. 85

Figure 4.20: Active information storage for BP and four EEG sub-bands. 87

Figure 4.21: The relationship between CRIB, MAP and EEG energy. (A) CRIB score and AMI between MAP and EEG energy (0.3-3Hz); (B) MAP and AMI; (C) CRIB score and MAP. Every point on the scatter plots represents 1 newborn. 88

Figure 4.22: The relationship between CRIB, MAP and EEG energy for the data during the first 24 hours of life only. 89

Figure 4.23: The relationship between the CRIB score and interaction between MAP and EEG (0.3-3 Hz) energy. (A) CRIB score vs TE from MAP to EEG (0.3-3 Hz); (B) CRIB score vs TE from EEG (0.3-3 Hz) to MAP. 89

Figure 4.24: The distribution of TE values for real data and randomly permuted surrogates. TE is quantified from MAP to EEG (0.3-3 Hz) for real data (yellow) and corresponding randomly permuted surrogates (black). TE from EEG (0.3-3Hz) to MAP for real data (green) and its randomly permuted surrogates (red). 90

Figure 4.25: Visualisation of the MAP along with the interaction of the MAP and EEG power (0.3-3 Hz) (represented as a cumulative average of AMI) for the preterm neonate (28 weeks GA). Missing values on the traces correspond to the eliminated regions affected by artefacts. 91

Figure 4.26: Schematic representation of the coupling between EEG and MAP using the autoregulation curve. The plateau of the autoregulation curve is used as a benchmark of normal brain function. The MAP and EEG are represented by the sequence of states denoted with letters. A higher level of interaction is implied by a longer overlap in the sequences (same patterns). When the MAP falls below an unknown threshold, the dynamic interaction between EEG and MAP changes. A higher risk of mortality which is represented with higher CRIB scores is shown to be associated with a smaller interaction between EEG and MAP and a stronger directionality of this interaction (from EEG to BP). 94

Figure 4.27: An association between MAP and GA. 95

Figure 4.28: An association between EEG power (15-30 Hz) and MAP. 95

Figure 5.1: The timing of illness scores assigned to an infant during the course in the neonatal intensive care unit (NICU) through to the neurodevelopmental follow-up at 2 years of age. In-patient risks include major neonatal complications: IVH (intraventricular haemorrhage), cystic periventricular leukomalacia, necrotizing enterocolitis, infection (sepsis), retinopathy of prematurity (ROP). The diagram is adapted from Lloyd et al. (Lloyd et al., 2016). 99

Figure 5.2: Schematic representation of duration and temporal location of recordings. Each recording is represented with respect to the time of birth (TOB) for each neonate.....100

Figure 5.3: One minute of raw ECG and eight-channel EEG and ten minutes of mean arterial pressure (MAP) recordings (GA=26 weeks).....101

Figure 5.4: Class specific histograms of the thirteen features extracted from all HRV epochs of 23 preterm neonates. No distinctive separation can be observed between the two classes.....104

Figure 5.5: PDF for the RMSSD feature. Original subset (a) contains RMSSD feature values from the complete recordings. Normal (b) and hypotensive (c) subsets represent RMSSD feature extracted during episodes of normal BP (MAP > GA+4) and during hypotensive events (MAP ≤ GA+4). PDFs are demonstrated for all 23 subjects (All GA), GA>28 subset (6 subjects) and GA≤28 subset (17 subjects).....106

Figure 5.6: Values of HRV features extracted from All epochs dataset for healthy and unhealthy preterm neonates. Boxplot analyses show the median, 25th and 75th percentiles, and the outliers. '**' represent statistically significant differences between groups with p<0.001 using Mann-Whitney U test. The predictive power of the features quantified by AUC is presented in Table 5.3 (All epochs, Set 3).....106

Figure 5.7: Principal component analysis (PCA) of the dataset comprised of all epoch (a) and epoch during hypotensive episodes (MAP ≤ GA+4) (b). In this study, PCA is used as a tool for exploratory feature analysis which is aimed at checking the discriminative power of the HRV feature set with respect to the short-term outcome of the preterm neonate.....107

Figure 5.8: An example of the association of the MAP with MeanRR and RMSSD features for one healthy (GA=29 weeks) and one unhealthy (GA=23 weeks) neonate. Shaded regions correspond to 95% CI. The correlation between MeanRR and MAP for the healthy and sick neonates is r = 0.32 and r = 0.15 correspondingly; and between RMSSD and MAP: r = 0.2 and r = 0.14. 109

Figure 5.9: Class specific histograms of the EEG features extracted from all epochs of 25 preterm neonates; no clear separation can be observed between healthy and unhealthy neonates. 109

Figure 5.10: AUC of three EEG and HRV features for various thresholds and the **All epochs** dataset. 111

Figure 5.11: The overall diagram of the multimodal short-term outcome prediction using a boosted decision tree classifier. Mean arterial pressure (MAP) threshold, is used as a data selection technique.....112

Figure 5.12: An example of the nonlinear decision boundary constructed with boosted decision tree classifier using two HRV features with (a) 50 trees, (b) 300 trees and (c) 1000 decision tees. It can be seen that the decision boundary constructed with 50 trees (a) differs from the one constructed with 100 trees. At the same time, no apparent distinction can be observed between 300 (b) and 1000 (c) trees. 113

Figure 5.13: A general diagram of the ensemble classification framework.....114

Figure 5.14: Visualisation of gradient boosting predictions using simulated data. 116

Figure 5.15: The visualization of an individual decision tree constructed using XGBoost; f12, f0, and f3 correspond to the HRV features: RMSSD, VLF, and LF/HF respectively. Every decision tree is trained to find out whether preterm has a good or poor outcome. Prediction score is assigned to every leaf – prediction for data points which fall into the given leaf. The final prediction is then obtained as the sum of scores predicted by each of the trees. 121

Figure 5.16: A diagram of LOO subject independent performance assessment and 5 times 2-fold CV model selection routines. 122

Figure 5.17: Mean of the feature importance (gain) reported by the boosted decision tree classifier trained on 13 HRV features extracted during episodes of low BP (MAP≤GA+4).....125

Figure 5.18: Representation of the importance (gain) of the top ten two-feature interactions for the short-term outcome prediction. The system is trained on the 13 HRV features extracted during the episodes of low BP (MAP≤GA+4). 125

Figure 5.19: Mean of the feature importance (gain) reported by the boosted decision tree classifier trained on all epoch of the EEG and BP features. 125

Figure 5.20: Representation of the importance (gain) of the top ten two-feature interactions for the short-term outcome prediction. The system is trained on all EEG and BP features extracted from the full recordings. 126

Figure 5.21: The density of selected tensors of three main hyperparameters obtained during the LOO routine for the HRV-based systems for Hypotensive events (MAP≤GA+4). The most frequently selected parameters are Subsample=0.9, Colsample=0.3 and Depth=4. The projections of the parameters are represented with dash lines. 126

Figure 5.22: Feature selection methods. 127

Figure 5.23: Mean of the feature importance (gain) reported by two boosted decision tree classifiers trained on the (a) HRV features extracted during episodes of low BP (MAP≤GA+4), and (b) EEG and MAP features extracted from all epoch. An additional randomly generated feature 'rand' is incorporated. The obtained results indicate that all EEG and HRV features bear some predictive information when compared to the random one. 128

Figure 5.24: A diagram of LOO subject independent performance assessment and 5 times 2-fold CV feature and model selection routines. 130

Figure 5.25: An example of the system output as a continuous probabilistic trace obtained during 10 hours for one healthy (GA=28 weeks, blue solid line) and one unhealthy (GA=23 weeks, red dashed line) patients. The system is trained and evaluated on the **All epochs** dataset. 131

Figure 5.26: Comparison of the probabilistic traces for an unhealthy neonate (GA=23 weeks) obtained from the two models trained on HRV features extracted from either the All epochs (b) or Hypotensive events (MAP<GA+4) (c) datasets. The model trained on the All epochs (b) is represented by instantaneous (red solid thin line) and cumulative (green dashed bold line) probabilistic values. The Hypotensive events model (c) is represented by the instantaneous probabilistic values (solid blue thin line) and the cumulative average of

prediction (solid orange bold line). An average of the morbidity prediction for both models is 0.85 and 0.95 correspondingly. 132

Figure 6.1: The timing of the Bayley scales along with other health scores assigned to an infant in the neonatal intensive care unit (NICU). It can be seen that from the time of the EEG recording (35 weeks GA) to the neurodevelopmental assessment (2 years of age) the neonate is exposed to various risks and confounding factors which are not taken into account but could potentially affect the developmental process. 137

Figure 6.2: An example of the 25-second long trace of burst annotations for one channel of raw EEG (T4-C4, fs=64 Hz) from one healthy neonate. Bursts are annotated with ones, whereas IBIs are marked with zeros. The mean IBI duration for the given EEG trace is ~ 4 seconds, which is within a normal range for the preterm with the GA of 35 weeks (Pavlidis et al., 2017). 140

Figure 6.3: Boxplots of the clinical features with respect to 2-year outcome based on the composite of three scales and language scale. 143

Figure 6.4: An example of CNN with fully connected layers. 147

Figure 6.5: An example of fully convolutional network architecture. 148

Figure 6.6: Schematic diagram of the fully convolutional neural network (16 filters) used for the 2-year language outcome prediction. 149

Figure 6.7: Schematic representation of the residual block within the network from Figure 6.6. 150

Figure 6.8: An example of the training AUC for each LOO iteration (grey), and its corresponding mean AUC (blue) for the network with 16 filters. 151

LIST OF TABLES

Table 2.1: Main ECG artefacts with corresponding affected frequency bands.....26

Table 3.1: EEG features. 37

Table 3.2: The frequency bands for the detection of the QRS complexes previously proposed in the literature (Elgendi et al., 2010). 38

Table 3.3: Frequency- and time-domain features extracted from the ECG. 45

Table 4.1: Clinical information represented as median (IQR). 73

Table 4.2: Correlation coefficient and 95% CI (in brackets) between CRIB score and coupling measures (correlation, coherence, AMI, and TE) between MAP and four EEG sub-band powers.....87

Table 5.1: Clinical information for preterms in the dataset. 100

Table 5.2: Frequency- and time-domain features extracted from EEG, ECG, and BP.....103

Table 5.3: The predictive power of the HRV features measured using AUC. 'All epochs' represents complete recordings. 'Hypotensive events' represents only epochs under the specific MAP threshold (GA, GA+2 or GA+4) for the same set of babies. Δ represents a change (an increase or decrease) of the AUC. 104

Table 5.4: The predictive power of the EEG features measured using AUC. 'All epochs' represent complete recordings. 'Hypotensive event' represents only epoch under the specific MAP threshold (GA, GA+2 or GA+4) for the same set of babies. Δ represents a change (an increase or decrease) of the AUC. 110

Table 5.5: AUC for short-term outcome prediction using various combination of HRV, EEG and BP features (MAP≤GA+4). AUCs of the best performing HRV- and EEG-based systems are in bold. 124

Table 6.1: Details of the Bayley-III assessment and a corresponding binary outcome. Sick: 24%, healthy: 76% for composite scale; sick: 19%, healthy: 81% for the language scale. The outcome is defined as abnormal (in bold) if 1) the value any of the three Bayley subscales is less than 85 for the composite outcome; 2) abnormal language outcome if the value of the language scale alone is less than 85. 138

Table 6.2: Frequency- and time-domain EEG features.....141

Table 6.3: The predictive power of the EEG features with respect to 2-year composite (3 scores) and language outcomes is measured using the AUC. Results are provided for the dataset of 37 preterm neonates. The highest AUCs are in bold. The AUC was quantified for various epoch lengths (1 min, 10 min, 20 min, and the whole recording). 142

Table 6.4: AUC values for the long-term outcome prediction. The outcome is represented as a composite of 3 Bayley scales and single language scale. AUCs for the best performing EEG-based system are in bold. 145

Table 6.5: AUC for the long-term language outcome prediction using the proposed CNN architecture. The AUCs obtained using the averaged predictions generated by the network with different initialisations are in bold. 151

LIST OF ACRONYMS

AF	Allan Factor
AMI	Adjusted mutual information
ApEn	Approximate entropy
BP	Blood pressure
CA	Cerebral autoregulation
CBF	Cerebral blood flow
CCS	Clinical course score
CI	Confidence interval
CNN	Convolutional neural network
CRIB	Clinical risk index for babies
CV	Cross validation
DFT	Discrete Fourier transform
DL	Deep learning
DP	Diastolic blood pressure
DTI	Diffusion tensor imaging
ECG	Electrocardiogram
EEG	Electroencephalogram
FC	Fully connected
GA	Gestational age
GAP	Global average pooling
GMM	Gaussian mixture model
HF	High frequency
HIE	Hypoxic Ischemic Encephalopathy
HR	Heart rate
HRV	Heart rate variability
IBI	Inter-burst interval
IQR	Interquartile range
IVH	Intraventricular hemorrhage
LF	Low frequency
LOO	Leave-one-patient-out
MAP	Mean arterial blood pressure
MI	Mutual information
ML	Machine learning

NICU	Neonatal intensive care unit
NIRS	Near-infrared spectroscopy
NN	Neural network
PCA	Principal component analysis
PDF	Probability density function
PMA	Post menstrual age
PSD	Power spectral density
ReLU	Rectified linear unit
RMSSD	Root mean square of the successive differences
RP	Relative power
SE	Spectral entropy
SGD	Stochastic gradient descent
SP	Systolic blood pressure
SVM	Support vector machine
TE	Transfer entropy
VLBW	Very low birth weight
VLF	Very low frequency

Chapter 1: Introduction

1.1 Premature birth

Prematurity is a leading cause of death in children under the age of five (Dempsey and Barrington, 2007) with more than one million children dying each year due to the complications of preterm birth. Significant advances in neonatal intensive care over the last number of decades have led to an increased survival rate for extremely preterm neonates. A major aim of the modern neonatal intensive care is to prevent brain injury in preterm neonates, however, the burden of disability remains high. A technique that could accurately interpret physiological signals, such as electrocortical brain activity, blood pressure (BP), and heart rate (HR) in the preterm and identify those neonates at risk of adverse outcome, would greatly aid the management of this vulnerable group in the neonatal intensive care unit (NICU).

Physiological signals are a valuable source of information that can provide an insight into the function of the human body. These signals are considered to be the main health indicators which can assist clinicians in the decision making related to the diagnostics and treatment of various diseases. Physiological signals are obtained from the sensors placed on the different parts of the patient's body. In order to use the recorded physiological data for healthcare applications, foremost, it is necessary to establish which information is relevant to the problem of interest. This can be a challenging undertaking (Miotto et al., 2018), as it requires domain-specific knowledge, which might take years of training to obtain. This is further complicated by fatigue-related mistakes as well as inter and intra-subject variability (Morrell and Morrell, 1966). These are the main constraints which contribute to the difficulty of clinical decision making.

Hypotension, or low blood pressure (BP), is a common problem in preterm babies, particularly in the first 72 hours after delivery. It may cause decreased cerebral perfusion, resulting in impaired oxygen delivery to the brain (Victor et al., 2006b). The criteria which define hypotension have not been clearly set (Dempsey and Barrington, 2007) and the decision on when and whether it should be treated remains disputed, resulting in considerable variability in clinical practice (Dempsey and Barrington, 2006), (Laughon et al., 2007). Treatment often involves administration of volume expanders and inotropes with dopamine as

a first-line agent when the mean arterial pressure (MAP, in mmHg) falls below the gestational age (GA) in weeks (Dempsey, 2015). This approach, however, is not supported by any robust scientific evidence (Development of audit measures and guidelines for good practice in the management of neonatal respiratory distress syndrome, 1992). At the same time, excessive intervention in order to treat hypotension in preterm infants has been associated with adverse outcomes, including brain injury (Synnes et al., 2001). Often, preterm infants with low BP have no biochemical or clinical signs of shock and thus may not require any treatment. In this case, a "permissive hypotension" approach, which implies careful observation without intervention may well be appropriate (Dempsey et al., 2009).

Due to the absence of any firm guideline on the assessment of hypotension in preterm neonates, it is important that not only BP but also other physiological signals including electroencephalogram (EEG), near-infrared spectroscopy (NIRS) and electrocardiogram (ECG) are monitored. EEG and NIRS are commonly used technologies to assess the 'brain health' of a newborn. EEG provides information about electrical cortical activity in the neonate (Laughon et al., 2007); NIRS allows continuous monitoring of cerebral oxygen saturation (tissue oxygenation) in the brain (Sood et al., 2015). Both methods are non-invasive and provide a real-time insight into the brain function. ECG, on the other hand, allows measuring the electrical activity of the heart.

Modalities and interaction

While all these signals can be monitored, little is known about their relationship in preterm infants as the extraction and investigation of the complex measures of signal interaction and signal dynamics have not yet been fully explored. Deciding when and whether to treat hypotension relies on our understanding of this relation between BP, oxygenation and brain activity.

Detection of relationships and the quantification of interactions between physiological systems can be carried out in different ways, with classical linear methods, coherence, and correlation being the most common. Given two time series, the latter measures the level of linear coupling in the time domain, whereas the former quantifies the interaction between signals in the frequency domain. The physiological signals, EEG in particular, are known to result from complex nonlinear processes, and the relationship between such signals is likely to be non-stationary, where short periods of coupling may occur at different points of time. As a result, in order to measure the linear, stationary association between such non-stationary signals a sliding window approach is applied, with the assumption that the relationship is stationary within the window (Pfurtscheller et al., 2012). While solving the problem of non-stationarity, this approach fails to detect nonlinear coupling between the signals. Due to the likely complex

relation between physiological systems, nonlinear methods of interaction based on information theory are usually applied (Pikovsky et al., 2003). According to Netoff et al., (2006) nonlinear measures are very sensitive to noise and linear methods sometimes present better properties in this sense (Pereda et al., 2005). However, both linear and nonlinear approaches assess different aspects of the interdependence between the signals and together may provide a more comprehensive picture of the analysed data.

The detection of the presence of a dominant direction for the coupling between physiological systems can provide an insight into their mutual interdependency. In order to capture the causal relationship, a time lag can be introduced in either one of the sequences. Granger causality is a statistical test (based on the linear regression modelling), which allows for the investigation of causal influence between two signals (Granger, 1969). Transfer entropy is a nonlinear measure which quantifies the exchange of information between two sequences (Schreiber, 2000). Both, TE and Granger causality were previously used in the field of neuroscience for establishing the underlying directed information structure between brain regions (Coben and Mohammad-Rezazadeh, 2015; Lizier et al., 2011).

Several studies have tried to establish the relationship between EEG activity and BP. D. Shah et al., (2013) identified that BP and EEG energy were associated with blood flow in the superior vena cava in the first 12h of life. However, West et al., (2006) found no association between superior vena cava flow and EEG energy. Increased oxygen extraction has been related to spontaneous activity transients observed in the EEG during the first 6 hours of life (Tataranno et al., 2015). The levels of BP which result in abnormal cerebral activity, as quantified by EEG spectral features (obtained from 4 bipolar channels) and peripheral blood flow measured with NIRS were studied in 35 very low birth weight infants in (Victor et al., 2006b), where it was reported that a low BP (below 23 mmHg) caused an increase in EEG discontinuity and a decreased relative power in the delta band (0.5-3.5 Hz). Changes in preterm EEG spectral power with maturation were also observed in a study by Niemarkt et al., (2011). Most of these studies were performed on short EEG recordings utilizing only a single summary measure of the BP and EEG computed from the whole recording.

ECG is yet another important physiological measure routinely recorded in preterm infants. It is usually quantified by the time variation between successive heartbeats (heart rate variability, HRV). HRV provides a non-invasive assessment of both the sympathetic and parasympathetic control of the heart rate (Akselrod et al., 1981). It is uniquely suitable for investigation of the influence of the immature autonomic system on cardio-respiratory control in preterm neonates. The information extracted from both the EEG and ECG recordings has been extensively used for assessing various aspects of newborn health. Promising results have been obtained using

the automated computer-based outcome prediction in full-term neonates based on a combination of multimodal features including HRV and EEG (Temko et al., 2015). In term neonates, a significant association between HRV, the severity of hypoxic-ischemic brain injury and long-term neurodevelopmental outcome at two years of age was reported for 61 full-term neonates (Goulding et al., 2015). For the preterm population, different physiological modalities have also shown promise in predicting the neurodevelopmental outcome (Lloyd et al., 2016; Périvier et al., 2016). A quantitative analysis of EEG, heart rate (HR), peripheral oxygen saturation and other clinical features from 43 preterm infants was conducted by Lloyd et al., (2016). Logistic regression analysis of all combined measures showed the potential for the prediction of both mortality and 2-year outcome. A correlation between low-frequency oscillation of HRV and BP for preterm neonates has also been reported by Rassi et al., (2005). Another study on 92 preterm neonates revealed a significant association between neonatal HRV and respiratory distress syndrome and proposed the use of HRV as an indicator of morbidity and mortality in preterm infants (Cabal et al., 1980).

1.2 Decision support in NICU

Machine learning (ML) is a field of study that focuses on how computers learn from the data without being explicitly programmed. It is subdivided into two main categories: supervised and unsupervised learning. Supervised ML is most frequently used in the clinical setting as it allows for the estimation of risk. A common example of this technique would be an automated detection of the heart disease based on the interpretation of the ECG signals, where the computer approximates what a trained cardiologist is doing with high accuracy (Deo, 2015). Unsupervised learning, on the other hand, does not perform output prediction, but rather looks for patterns occurring within the data. This method was previously used for the discovery of new genomic features, where clusters of genes shared similarity across tumor samples obtained from different cancer types (Cheng et al., 2013b). This information allowed to develop a prognostic model for breast cancer survival (Cheng et al., 2013a).

Nowadays, **computer-based analysis** of signals is widely used for healthcare applications (Faust et al., 2012). These systems allow for accurate and objective decisions which are not affected by human-related mistakes; as a result, a patient is able to receive faster and more effective care. The advances in data acquisition systems have allowed for the availability of potentially large datasets of physiological signals. This has created an exceptional opportunity for the development of an individualised medical treatment.

The neonatal population is especially vulnerable and prone to various health complications. A delay in the problem identification, as well as delays in the providing of the correct

management, may be fatal for this vulnerable group. This has created a special need for computer-based medical devices in the NICU which can assist clinicians in decision-related tasks. Preventing health problems at the early stage is known to be cheaper and more efficient as it allows for the identification of at-risk patients in a timely manner (Edmond and Zaidi, 2010). All preterm infants are at a high risk of re-hospitalization and further health complications, which affects the life of parents (Cronin et al., 1995). Additional emotional, physical and financial burdens have shown to negatively influence the quality of families' life (Lakshmanan et al., 2017). As a result, improved timing and quality of the treatment can potentially have a far-reaching effect on the family, health service and society at large.

1.2.1 Computer-based analysis: classical approach

The classical approach of computer-based analysis is comprised of the following steps: data pre-processing, feature engineering, feature extraction, and modelling. The model can be constructed in different ways. The artificial neural network (NN) along with the backpropagation algorithm was initially proposed by Werbos, (1975). It created an opportunity for efficient training of the multi-layer network, which allows for learning complex nonlinear representations of the input feature space. This was followed by yet another popular ML algorithm based on the decision trees, known as classification and regression trees (CART) analysis (Breiman et al., 1984). Significant improvements in the classification accuracy of decision trees were achieved with ensemble techniques such as random forest, where trees are constructed on a randomly selected subset of features (Breiman, 2001). Bagging was introduced to construct trees by repeatedly sampling the data with replacement followed by the voting scheme to generate a final prediction (Breiman, 1996). Boosting was proposed as a method for the creation of an accurate classifier from a set of weak classifiers (decision trees). This can be constructed in a stage-wise manner with adaptive boosting, known as AdaBoost (Freund and Schapire, 1996) and gradient boosting (Friedman, 1999) for example. Constant enhancements in the field of decision trees modelling have allowed this technique to be widely used nowadays, with the recently developed algorithm gradient boosting with regularisation (XGBoost) (Chen and Guestrin, 2016), which is widely recognized in a number of ML challenges (e.g. Kaggle competitions). Another important breakthrough in ML was related to the emergence of support vector machines (SVM) (Cortes and Vapnik, 1995). This technique allows for the construction of the classifier by projecting the original feature set onto the higher dimension in order to discriminate the data. However, in contrast to "black box" modelling techniques such as NN and SVM, the main advantage of the tree-based classifier lies in its interpretability, which is essential in the medical field.

One of the examples of the classical computer-based analysis of the physiological signals in newborns is aimed at early detection of a seizure onset. Neonatal seizure is one of the numerous examples where medical teams struggle to recognise the problem by human eye. This is mainly due to the insufficient number of experts being able to interpret the physiological signals, EEG in particular. The presence of seizures has been associated with an increased brain injury, which may consequently lead to further complications (Rennie and Boylan, 2007). The problem of seizure detection was successfully addressed with an SVM-based algorithm using 55 hand-crafted features extracted from the EEG recordings (Temko et al., 2011). The output of the SVM classifier generates a probability of a seizure for every 8 seconds of EEG. The developed system allows for an acute interpretation of the brain signals and alerts medical staff when abnormal brain activity is detected. The algorithm for the neonatal seizure detection (ANSeR) successfully passed the clinical trials and proved to be useful in a real clinical setting ("ANSeR- The Algorithm for Neonatal Seizure Recognition Study - ClinicalTrials.gov"; Mathieson et al., 2016). The availability of such an AI-powered monitor allows clinicians to detect seizures and provide necessary treatment in a timely manner.

The heart rate observation monitor (HeRO) (Hicks and Fairchild, 2013) is yet another example of a computer-based system applied to physiological signals in the NICU. Preterm neonates are prone to various diseases and complications; this includes sepsis, which is known to be an important risk factor for a prolonged stay at the hospital and even death (Lake et al., 2002; Stoll et al., 2002). HeRO is an early alarm system aimed at early sepsis detection. The system analyses the heart activity of the neonate based on three heart rate characteristics, namely, the standard deviation of the beat-to-beat intervals, the sample asymmetry which quantifies both acceleration and deceleration of the heart rate, and the sample entropy to measure the irregularity of the HRV. The final algorithm is based on logistic regression, which generates a clinical score that quantifies the risk of sepsis in the next 24 hours (Fairchild and Aschner, 2012). Results of a large randomised clinical trial have reported a reduction of 22% in mortality when providing a sepsis score to the medical staff. The system is now used in many NICUs in the USA and is also approved for use in Europe.

Classical feature-based algorithms have shown to be quite successful and were used for a number of tasks related to the various problems of the adult and neonatal health. This approach, however, is known to rely on the prior knowledge of physiological signals, and therefore it is prone to human bias. It is worth mentioning, that physiological signals are known to have a complex structure and the extraction of the relevant features requires a good understanding of the underlying physiological processes. At the same time, the performance of the feature-based algorithms is highly dependent on the feature set itself. Consequently, in the situation when a-

priori extracted features are not able to fully capture all the important information present within a signal, the performance of the system can be potentially affected.

1.2.2 Computer-based analysis: end-to-end learning

An alternative approach of the computer-based system is based on the end-to-end learning. A new wave of deep learning (DL) algorithms allows for overcoming the problems of the feature extraction and feature selection. Such algorithms do not require any hand-crafted features as they enable learning of the relevant features from the input data by performing the end-to-end DL (O'Shea et al., 2017). This technique replaces the multiple steps of the feature extraction, feature selection, and classification with a single NN. Scale of NN drives DL progress and the performance of DL techniques is highly dependent on the amount of available labeled data examples. New sophisticated sensors connected through the internet of things are now used in modern medical equipment. Being able to digitise almost every aspect of the human body, along with advances in ML, has led to the era of big data in the healthcare field. According to the report ("Becoming A Data-Driven CEO | Domo,"), over 2.5 quintillion bytes of data are being created every day, and deep NN has shown the capability of taking advantage of it. More data allows for the construction of deeper models without overfitting. However, when the amount of the available training data is limited, the traditional algorithms which rely on the feature engineering may still outperform these new deep NN paradigms.

One of the examples of the end-to-end learning is convolutional neural network (CNN), a class of deep NN that can learn relevant features using the end-to-end hierarchical representation of the data. CNNs were originally developed for image processing and are well-known for their great success in solving computer vision problems (Hinton et al., 2012; Krizhevsky et al., 2012), audio, music and speech processing tasks (Abdel-Hamid et al., 2014; Lee et al., 2009; Schlüter and Böck, 2014). In the field of physiological signals processing, EEG spectrograms were used for the task of epilepsy prediction in adults (Korshunova et al., 2018). O'Shea et al., (2017) proposed a CNN which has shown to be comparable with the state-of-the-art SVM-based algorithm for neonatal seizure detection, where instead of using hand-crafted features, the authors applied fully convolutional architecture, which generates the probability of seizure from raw multichannel EEG data.

While there has been a considerable success of development and deployment of ML in NICU for the term neonates (O'Shea et al., 2017), preterm infants remain to be at a higher risk of complications, which may lead to both short-term and long-term adverse health outcomes including neuromotor, cognitive, hearing, and visual problems (Doyle et al., 2003; Kistner et al., 2000; Siewert-Delle and Ljungman, 1998). These risks are even higher the earlier a baby

is born. This thesis aims at filling this gap by developing intelligent monitoring of preterm infants in NICUs.

1.3 Aims and scopes of the thesis: Intelligent monitoring of preterm neonates

Multimodal signal analysis is required to provide an insight into the physiology of the preterm neonate and to better understand the interrelation between the brain (EEG), heart activity (HRV) and BP in the context of the infant's health status. Physiological signals are rarely recorded simultaneously, which poses a significant challenge in the field of neonatal health research. This is due to the fact that preterm neonates are very vulnerable and most of the time it is difficult to get permission from the neonatologists for any kind of intervention. The physiological signals used in this thesis were provided by the Infant research centre, a world-leading research centre for fetal and neonatal translational research. Having such invaluable access to the database of long unedited multimodal recording from preterm neonates recorded soon after birth has allowed us to conduct research which can potentially contribute towards more efficient management of the preterm population.

After birth, preterms are admitted to the NICU, where in an intrauterine-like environment they get breathing support, feeding, and treatment for infections (Figure 1.1). Infants will stay in the NICU until they reach a stable health condition, including an appropriate weight, stable temperature and absence of diseases (sepsis, apnea, etc.). As a result, it might take months before they are discharged home.

Figure 1.1: A preterm neonate being monitored in the NICU at Cork University Maternity Hospital.

The health status of the preterm is usually evaluated at various stages of life using different clinical measures. In this thesis, the assessment of the well-being of the preterm was performed

at three different stages as schematically represented in Figure 1.2. This starts from the admission to NICU, where at the first 12 hours of life the health status is quantified by the clinical risk index for babies (CRIB II) (Parry et al., 2003). This is followed by the assessment of the preterm's well-being at discharge using the clinical course score (CCS) (Lloyd et al., 2016). Two years later it is possible to perform the long-term follow-up, where the neurodevelopmental outcome is assessed using the Bayley III scales (Bayley, 2006) as shown in Figure 1.2.

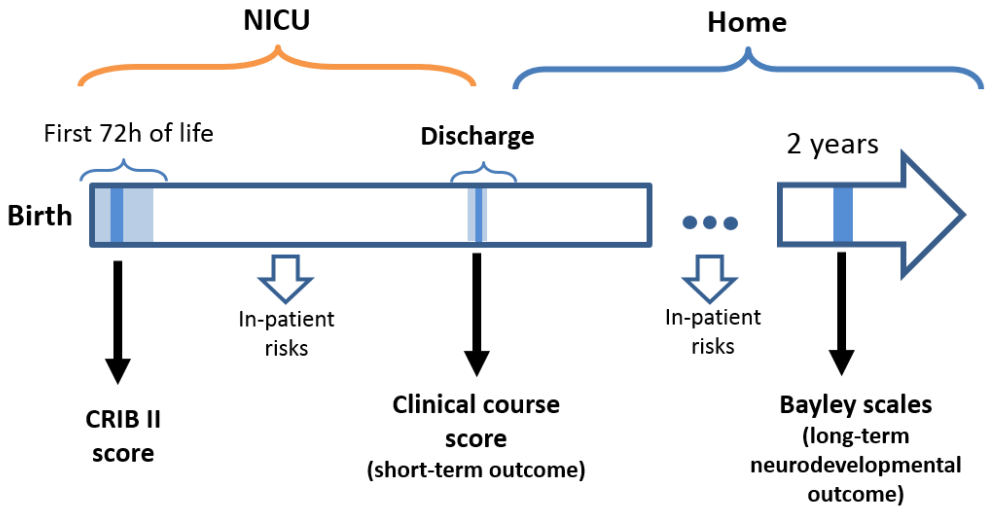


Figure 1.2: The timeline of an infant's stay in the NICU through to the long-term neurodevelopmental follow-up at 2 years of age.

The main aim of this thesis is to develop decision support systems for intelligent monitoring of the preterms in NICUs. In order to achieve this the following objectives were defined in this thesis and researched:

Objective 1 of this thesis is to investigate the coupling between various physiological signals recorded from the premature neonates in NICUs and to explore how this interaction is affected by the early health status of the preterm as assessed by the CRIB score.

While the area of hypotension management in preterms remains untested, the problem of low BP is a very common complication which usually occurs during the first 3 days of life. According to the current concept of health, a coupling between organs or in some cases the absence of coupling may be considered as a sign of health (Godin and Buchman, 1996). The level of coupling between two systems, brain activity (quantified with 8 bipolar EEG channels), and BP (recorded invasively), was quantified using classical linear methods such as coherence and correlation, and a nonlinear measure of mutual information. In order to provide a better insight into their mutual interdependency, the presense of a causal relationship was quantified using transfer entropy. Hypothesising that differences in coupling between brain activity and BP are indicative of the preterm's wellbeing, the association between this

dynamic coupling and an illness risk score was tested. We anticipate that computing and visualising the measure of interaction between BP and cerebral activity would be feasible as part of a cotside monitor in order to provide real-time decision support for more efficient management of hypotension in preterm neonates.

Objective 2 of this thesis is to investigate the relationship between different modalities (EEG, ECG, and BP) and the short-term health outcome of the preterm quantified with the CCS at a discharge (Figure 1.2).

Both EEG and ECG have defined ranges of characteristics which correspond to the normal well-being of the neonate (Pavlidis et al., 2017; Selig et al., 2011). This information allows for the proper interpretation of the current health status as well as the potential to predict the future well-being of the neonate. While several studies have identified an association between HRV, EEG and neonatal health outcomes in both term and preterm infants, there is still a lack of understanding of this relationship in the context of low BP episodes in preterm infants. In particular, the predictive capability of HRV and EEG for the estimation of short-term health outcome is assessed and the predictive power of both HRV and EEG features during the episodes of low BP is studied. This concept has been incorporated while combining multimodal data of ECG, EEG, and BP with an objective to develop an ML-based decision support tool for clinical prediction of the outcome at the discharge.

Objective 3 of this thesis is to investigate the association between the EEG signals recorded at discharge and the long-term outcome quantified using the Bayley scales of neurodevelopmental outcome assessed at 2 years of age.

Preterm neonates born earlier than 32 weeks GA are known to be at a higher risk of morbidity as compared to older preterms (Larroque et al., 2008). At the same time, preterm infants were reported to have a significantly higher disorganized behaviour (e.g. motoric, attentional, reactivity) as compared to full-term neonates assessed at two weeks after the due date (Als et al., 1988). Early interventions for preterm neonates may offer a positive influence on their motor and cognitive outcomes (Spittle et al., 2015). Therefore, necessary treatment initiated soon after birth can potentially improve the neurodevelopmental outcome of the preterm infant. Nowadays, accurate early recognition of infants with an increased risk of adverse long-term outcome poses many challenges. Being able to detect at-risk babies may allow them to benefit from specific treatments (Spittle et al., 2015). This can also help doctors to communicate with parents by providing a reliable prediction of possible risks and complications. The EEG grading system developed by Pavlidis et al., (2019) has shown to be indicative of the impaired long-term neurodevelopmental outcome at 2 years of age. The proposed system is based on manually extracted EEG characteristics used to quantify brain

function. The procedure of manual grading of the EEG is difficult, as it is a time-consuming process that requires domain-specific knowledge. In a real-world clinical setting, every minute counts and the ability to promptly make accurate decisions regarding treatment procedures is of great importance. In this context, the study aims to investigate the predictive capability of early EEG with respect to 2-years outcome using ML techniques.

1.4 Thesis layout

The development of an intelligent decision support tool for the accurate estimation of the current health status of the preterm neonate as well as the prediction of possible long-term complications is a multidisciplinary field of research which encapsulates knowledge of neuroscience, medicine, engineering, and computer science. Therefore, the work conducted in this thesis has a broad focus and the thesis layout is represented as follows:

Chapter 2 provides the necessary background information on neonatal health and complications (hypotension in particular) occurring soon after birth as well as the description of different physiological parameters being recorded, such as EEG, ECG, and BP. For readers, who have no clinical background, this information may be relatively new and therefore, it is important to clarify the problems medical staff face in a real-world clinical setting when dealing with an extremely vulnerable population of preterm neonates.

Chapter 3 provides a description of the technical methods used in this thesis. These include physiological signals processing, feature extraction techniques and measures of coupling used to assess the interrelation between signals. This chapter also provides a brief description of the main ML paradigms and a short overview of the main supervised ML techniques along with their application in the medical field. The measures used to assess the performance of models are also discussed in this chapter.

Chapter 4 presents a study that aims at measuring the coupling between brain activity as quantified by the EEG, and BP as quantified with the MAP. The level of coupling between these two physiological systems is then investigated in the context of the preterm health status as measured by the CRIB score. The level of coupling was estimated using classical linear methods such as correlation and coherence, and the nonlinear method of mutual information. The causality of interaction was measured using transfer entropy. The reliability of the obtained results was checked by testing an appropriate null hypothesis for every computed measure of interaction using surrogates. This is done in order to define whether a given empirical non-zero measurement of interaction is statistically different from zero.

Chapter 5 presents a multimodal decision support system based on the EEG, ECG and BP recordings, for the prediction of the short-term health outcome in preterm infants. The chapter is separated into two main parts. The first part studies the predictive power of the HRV and EEG for the estimation of short-term health outcome for full recordings and during episodes of low BP (hypotension). The second part represents the EEG- and HRV-based classifiers developed using boosted decision trees along with the theoretical background of the used technique. Additionally, various feature extraction techniques which were used to further boost the performance of the classifier are described. The chapter closes with a detailed description and discussion of the developed decision support tools and its potential application in a real-world clinical setting.

Chapter 6 presents a study that aims at developing an automated system for the long-term neurodevelopmental outcome prediction. First, the relevant EEG features are extracted. Their predictive power is then assessed with respect to the 2-year neurodevelopmental outcome. It is shown that some EEG characteristics are useful for the problem. The second part investigates the capability of the classical feature-based classifier and end-to-end ML technique to predict the long-term outcome of preterm neonate.

Chapter 7 is the concluding chapter, which summarises the main findings and contributions of the thesis and discusses potential future directions.

1.5 List of publications arising from this thesis

Journals:

1. **Semenova, O.**, Lightbody, G., O'Toole, J.M., Boylan, G., Dempsey, E., Temko, A., 2018. Coupling between mean blood pressure and EEG in preterm neonates is associated with reduced illness severity scores. PLOS ONE 13, e0199587. https://doi.org/10.1371/journal.pone.0199587
2. **Semenova, O.**, Carra, G., Lightbody, G., Boylan, G., Dempsey, E., Temko, A., Prediction of short-term health outcomes in preterm neonates from heart-rate variability and blood pressure using boosted decision trees. Comput. Methods Programs Biomed. 180, 104996. https://doi.org/10.1016/j.cmpb.2019.104996

Conference proceedings:

1. **Semenova, O.**, Lightbody, G., O'Toole, J.M., Boylan, G., Dempsey, E., Temko, A., 2017. Modelling interactions between blood pressure and brain activity in preterm neonates. Conf. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf. 2017, 3969–3972. https://doi.org/10.1109/EMBC.2017.8037725

2. **Semenova, O.**, Carra, G., Lightbody, G., Boylan, G., Dempsey, E., Temko, A., 2018. Heart Rate Variability during Periods of Low Blood Pressure as a Predictor of Short-Term Outcome in Preterms. 2018 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBC 5614–5517. https://doi.org/10.1109/EMBC.2018.8513600

Abstracts:

1. **Semenova O.**, O'Toole J., Boylan G., Dempsey E., Temko A., "A Method to Assess the Relation between Blood Pressure and EEG in Preterm infants", INFANT research seminar, June 2015.

2. **Semenova O.**, Lightbody G., O'Toole J., Boylan G., Dempsey E., Temko A., "Measuring Interaction between Blood Pressure and Brain Activity in Preterm Neonates", Conference on brain monitoring and neuroprotection in newborns, October 2017.
3. G. Carra, G. Lightbody, **O. Semenova**, E. Dempsey, A. Temko, "Heart Rate Variability during Periods of Low Blood Pressure in Preterm Neonates", Conference on brain monitoring and neuroprotection in newborns, October 2017.

Posters presented:

1. **Semenova O.**, O'Toole J., Boylan G., Dempsey E., Temko A., "A Method to Assess the Relation between Blood Pressure and EEG in Preterm infants", INFANT research seminar, June 2015.
2. **Semenova O.**, Lightbody G., O'Toole J., Boylan G., Dempsey E., Temko A., "Measuring Interaction between Blood Pressure and Brain Activity in Preterm Neonates", Conference on brain monitoring and neuroprotection in newborns, October 2017.
3. **Semenova O.**, Carra G., Lightbody G., Dempsey E., Boylan G., Temko A., "Prediction of short-term outcome in preterm neonates using boosted decision trees", The future of machine learning, February 2019

Chapter 2: Medical background – monitoring of preterm infants with hypotension

Continuous measurements of the various vital physiological parameters such as heart rate, pulse oximetry, respiration rate, and brain activity enable medical staff to always be informed about the changes in the health condition of the patient. These signals provide an insight into the functioning of various organs and systems and can help clinicians during the diagnostic process. In this chapter, the problem of hypotension in preterm neonates and its physiology is described along with other essential medical details of the related signals and systems. This information is important for the understanding of the experimental work presented in chapters 4, 5 and 6.

2.1 Cardiovascular system and blood pressure

Every year more than one in ten babies are born preterm and this number is rising ("WHO | Preterm birth," n.d.). Preterm birth occurs before the start of the 37 weeks gestational age (GA). Preterm babies are especially vulnerable and are at high risk of various complications. Hypotension or low BP is a recognised problem in preterm infants and it has been associated with various complications, which may lead to both short-term and long-term adverse health outcomes. Deciding when and whether to treat low BP is challenging. In order to properly address the problem of hypotension, it is necessary to step back and consider BP as part of the larger cardiovascular system. This will allow for a better understanding of the nature of hypotension through an appreciation of all the hemodynamic mechanisms taking place within the human body.

2.1.1 Neonatal cardiovascular system

The circulatory or cardiovascular system is an organ system that allows for blood circulation. This system makes it possible to circulate and transport oxygen, nutrients, carbon dioxide and other elements within the blood to and from body tissues in order to nourish them and maintain homeostasis. The cardiovascular system consists of the heart, blood and blood vessels. The blood vessels called arteries carry blood from the heart to lungs, where the blood enriches with oxygen. After passing through the body tissues blood vessels called veins carry deoxygenated

blood from the body tissues toward the heart again. The right heart then pumps the deoxygenated blood to the lungs and the left heart pumps oxygenated blood to the body through the blood vessels.

BP is the pressure of blood circulation on the walls of blood vessels. It is usually expressed in terms of the systolic (SP) and diastolic pressure (DP) and is measured in millimeters of mercury (mmHg). SP measures maximum arterial pressure during contraction of the left ventricle of the heart. DP is the minimum arterial pressure during the relaxation of the heart between beats and the dilation of the ventricles of the heart when they fill with blood. The optimal resting BP in an adult is approximately 120 mmHg systolic and 80 mmHg diastolic. BP can be influenced by various factors including emotional state, activity, the presence of certain diseases and many others. BP can vary from person to person as well as over life. More specifically, children and infants are known to have lower ranges of normal BP as compared to adults (Schwartz et al., 2002).

2.1.2 Blood pressure disorders: hypotension

In some conditions, BP may stay persistently low, high or erratic, which can lead to various health issues. The most common disorders of BP control are high BP, also called hypertension and low BP or hypotension. Hypertension in an adult is defined as SP higher than 140 mmHg or DP higher than 90 mmHg. In adults, if the SP is less than 90 mmHg or the DP is less than 60 mmHg, it is considered to be hypotension. Hypotension is usually accompanied by a number of symptoms such as dizziness, fainting, nausea, blurred vision, cold and pale skin, fatigue and others. Both disorders are well known in clinical practice and have established treatment guidelines for the adult population.

Hypotension is also a recognised problem in the population of preterm infants particularly during the first 72 hours after birth. However, unlike the adult population, where the ranges of normal BP are clearly identified, the definition of hypotension for premature infants is still uncertain. A number of studies have previously attempted to define the normal range of BP for newborn infants (Dempsey and Barrington, 2007). Results, however, have shown considerable variability due to a number of reasons including the small number of patients recruited in the study, an insufficient amount of data and others (Dempsey and Barrington, 2007). In clinical practice, the mean arterial pressure (MAP) measure is defined as the perfusion pressure of organs in the body and is used most commonly to determine the intervention criteria for preterm neonates. According to the recommendation of the Joint Working Group of the British Association of Perinatal Medicine, the MAP level (in mmHg) should not fall below the GA in weeks ("Report of a Joint Working Group of the British Association of Perinatal Medicine and the Research Unit of the Royal College of Physicians.,"

1992). This approach is not supported by any robust scientific evidence, however, it is the one which is the most commonly used in clinical practice to define hypotension in preterm infants.

In the study (Dempsey and Barrington, 2009) it was suggested that hypotension, or "unsafe BP" should be defined as the BP threshold under which there is a statistically increased risk of an adverse outcome. Another study (Barrington et al., 2002) attempted to answer this question on a large database of very low birth weight (VLBW) neonates. It identified an association between a statistically worse outcome (represented by the intraventricular haemorrhage (IVH) of grade 3 or 4) with a decreasing MAP threshold. The IVH was reported to increase from 21% to 31% when the hypotension definition was reduced from 20 to 15 mmHg for all patients with GA less than 28 weeks.

Ideally, the most appropriate definition of hypotension would be the one which defines a BP threshold, under which all treatment interventions will result in the improved outcome. However, it is unlikely that such an operational threshold can be represented with a single value of BP. Therefore, it was suggested (Dempsey and Barrington, 2009) that such a definition should be patient specific and depended on several factors, including birth weight, postnatal age, gestation as well as the possible cause of the hypotension itself.

2.1.3 Hypotension and autoregulation

The brain is the main command system of the human body which provides centralised control over other organs according to the input it gets from the sensory systems. It is one of the most metabolically active organs, which account for 20% of the resting energy of the body (Clarke and Sokoloff, 1999). High energetic and perfusion demands make the brain very vulnerable to ischaemic injury. Abnormally low BP can lead to an insufficient supply of oxygen to organs including brain, as a result, they may be temporarily or permanently damaged. A pathologic state which is characterised by inadequate oxygen delivery to tissues is known as shock. In a study by Osborn et al., (2004) authors reported very little or no correlation between low BP and systemic blood flow. This implies that low systemic perfusion and shock can occur with normal BP. At the same time, preterm infants with low BP may have no biochemical or clinical signs of shock and poor perfusion. In this case, the infant may not require any pharmacologic treatment and careful observation without intervention would be seen as appropriate (Dempsey et al., 2009). This approach was previously termed as "permissive hypotension".

The human body has several mechanisms to control BP and return it to the normal range after a disturbance. These include changing the blood volume in the blood vessels, changing the amount of blood that heart pumps to the body (cardiac output) and the diameter of the arteries. These mechanisms are controlled by the sympathetic nervous system, which regulates

processes within the body without any conscious effort. The BP is regulated by the autonomic nervous systems using baroreceptors, sensors located in the blood vessels. When a change in BP is detected, the nerves send this information to the brain to influence nervous and endocrine systems. More specifically, in order to stabilise low BP, a number of processes are activated: an increase in the HR in order to increase the amount of the pumped blood; constriction of veins in order to decrease the capacity to hold the blood; and constriction of arterioles in order to improve their resistance to blood flow.

The property of maintaining BP and organ perfusion at a sufficient level to preserve homeostasis is referred to as autoregulation. Cerebral autoregulation (CA) plays a dominant role in the homeostasis of cerebral blood flow (CBF). It allows for the maintenance of a stable CBF across a wide range of the MAP. In other words, CBF is stable when MAP is changing within a range of normal values.

2.1.4 Autoregulation in preterm neonates

The upper and lower permissible levels of BP where cerebral autoregulation is properly functioning are known for adults. However, for the preterm population, the permissible MAP bound is still uncertain. Greisen, (2005) reported that the lower permissible level of MAP for newborns starts at 30 mmHg, with no upper limit specified. Figure 2.1 represents the autoregulation process, where the plateau on the curve corresponds to a functioning CA. The slope of this part is near zero. When the MAP is outside of this range – autoregulation does not function anymore and cerebral blood flow (CBF) becomes pressure passive.

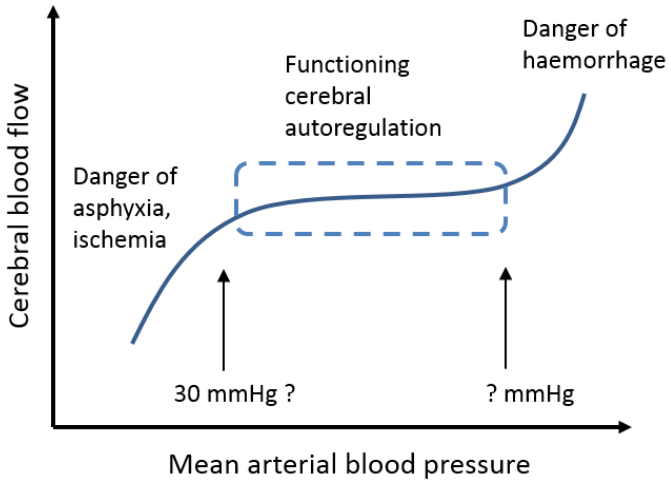


Figure 2.1: Schematic representation of cerebral autoregulation in preterm neonates. The plateau on the curve represents properly functioning autoregulation, where for a range of mean arterial pressure (MAP) the cerebral blood flow (CBF) does not change. This figure is adapted from (Greisen, 2005).

Accurate measurements for lower and upper thresholds for static autoregulation are still not fully determined and dynamic autoregulation has been reported not to operate in the immature brain of preterm babies at all. In the situation when the body is not able to compensate for the changes in BP, different organs including the brain may start to malfunction. Abnormally low MAP may lead to asphyxia and ischemia caused by insufficient brain perfusion. At the same time, a high MAP creates the danger of intracranial haemorrhage caused by the rupture of small capillaries. Monitoring MAP and its relation to CBF is of critical importance and it has become a very challenging task for clinicians who are dealing with preterm neonates.

2.1.5 Recording blood pressure

After birth, preterm neonates are admitted to NICUs. During this time it is essential to have access to various physiological signals including BP. Two techniques are currently available for measuring BP in newborns which can be characterised as either invasive or non-invasive methods. The non-invasive measurements are much simpler to acquire and require less expertise as compared to the invasive BP measurement. At the same time, non-invasive BP methods are less accurate and are more commonly used for routine examinations.

The most typical non-invasive methods are palpatory and auscultatory. The palpatory method allows to roughly estimate systolic BP by palpation. It is frequently used in emergency situations, however, this method does not allow for the estimation of diastolic BP. The auscultatory method uses mercury sphygmomanometers and stethoscope for the estimation of the BP level. A sphygmomanometer is composed of an inflatable cuff and a mercury or mechanical manometer. The manual sphygmomanometer is used with a stethoscope. The non-invasive methods also include automatic methods of BP measurement such as oscillometric monitor and arteriosonde (Doppler technique). The oscillometric method of BP measurement involves the observation of the pressure oscillations caused by blood flow in the sphygmomanometer cuff. This technique requires less skills as compared to the auscultatory methods and can be used for automated patient home monitoring. However, when recording BP in small and sick neonates, regular automatic inflation of the cuff may disturb the infant, especially those, whose conditions are not stable. Oscillometry is the most commonly used method in NICUs. The ultrasonic Doppler techniques estimate BP by detecting changes in blood flow during the external compression of the inflatable cuff. This technique is rarely used for neonates due to a number of drawbacks. More specifically, Dweck et al., (1974) reported a correlation coefficient of only 0.8 when comparing BP measurements obtained from the Doppler technique with the direct invasive measurement through the arterial catheter. Another study by Töllner et al., (1980) reported that Doppler technique may cause radial nerve palsy in premature infants due to the pressure of the transducer on the radial nerve.

The non-invasive techniques for BP measurement are known to be unreliable in small and sick infants and the invasive method is considered as the gold standard (Weindling, 1989). The invasive method involves direct measurement of arterial pressure by inserting a cannula needle in the umbilical artery. The main components of such a monitoring system are the measuring device, the transducer, and the monitor, where the waveform is displayed. The umbilical arterial catheter is connected to a BP transducer. The catheter is usually inserted using palpation or with the help of the ultrasound guidance. This type of BP monitoring allows for accurate measurements at very low levels of pressure during a long period of time. This is particularly suitable for the neonates who are admitted to NICUs and are suffering from hypotension and possible shock.

For the work conducted in this thesis, the gold-standard continuous invasive arterial BP monitoring was performed via an umbilical arterial catheter using the Philips Intellivue MP70 machine. All infants were nursed supine. The positioning of the tip of the umbilical catheter in the descending aorta was confirmed by chest radiograph.

2.1.6 Current therapies and outcomes

Unfortunately, it is still not possible neither to assess the adequacy of end-organ perfusion nor to diagnose shock in the preterm infant. As a result, the decision on when and whether low BP should be treated remains disputed, resulting in considerable variability in clinical practice (Dempsey and Barrington, 2006), (Laughon et al., 2007). Treatment often involves the administration of volume expanders followed by inotropic support with dopamine as a first-line agent.

It should be appreciated that an excessive intervention in preterm infants may be unnecessary, or possibly even harmful. Such excessive intervention was previously associated with morbidity and hearing loss (Vargo and Seri, 2011). Likewise, analysis of a large neonatal database (Canadian Neonatal Network) comprised of 5126 preterm infants with less than 33 weeks GA has demonstrated that the treatment of hypotension was associated with an increase in serious brain injury (Synnes et al., 2001). The authors suggested that the treatment of hypotension, rather than its presence, could be more harmful to the preterm.

Volume expanders

A volume expander is a well-known type of intravenous therapy aimed at stabilising circulatory haemodynamics by providing more volume to the circulatory system. This allows for the restoring of the proper oxygenation of body tissues. The usage of volume expansions is the initial approach employed by most clinicians in order to tackle hypotension by restoring proper perfusion of the body. However, there is no reliable evidence to support this approach

as many hypotensive infants have normal circulating blood volumes (Bauer et al., 1993; Kluckow and Evans, 1996). At the same time, volume expansion during the first days of life was shown to be associated with bronchopulmonary dysplasia (Van Marter et al., 1992, 1990). Goldberg et al., (1980) reported that the most common approach of rapid volume expansions used for BP stabilisation is associated with IVH. The study was conducted on an observational dataset of 214 infants (birth weight < 1400 gm; GA: 28.5±0.4 weeks). This may assume that bleeding could have happened when delicate blood vessels of the premature infant were not able to adapt to the rapid increase in BP after the infusion of volume expansions. Another study also reported an association between adverse neurologic outcome and colloid infusion for VLBW infants (Greenough et al., 2002), suggesting that this treatment should be used with caution in the perinatal period.

Inotropes

An inotrope is an agent that affects the force of muscle contraction. Dopamine is an inotropic heart drug that causes more intense contractions of the heart muscle, which consequently can increase the BP. Dopamine is known to be the most commonly used inotrope in clinical practice for the management of hypotension. At the same time, this drug was reported to be associated with a reduction in the systemic perfusion (Osborn et al., 2002; Zhang et al., 1999) and a poor long-term neurodevelopmental outcome in preterms (Filippi et al., 2007). Dobutamine is another inotrope which is extensively used in preterm neonates and is associated with an increase in superior vena cava flow and was reported to be unreliable in increasing BP (Osborn et al., 2002).

As we can see, the current clinical approaches of hypotension treatment may not be safe and may even be harmful for the infant. The "Hypotension in Preterm Infants" trial (HIP) (Dempsey et al., 2014) which was carried out by a multidisciplinary team of scientists, pharmacologists and neonatologist investigated whether a standard approach to the management of hypotension with volume expanders and dopamine, as opposed to a more observational approach with placebo, may result in improved short-term and long-term outcomes for preterm infants. The question of when and how hypotension should be treated still remains unanswered. The absence of a clear clinical diagnosis of shock in preterm neonates means that many infants may be exposed to unnecessary treatment based solely on whether their MAP was lower than the GA (weeks) threshold, which is not supported by any robust scientific evidence. Since excessive intervention in order to treat hypotension in preterm infants has been associated with adverse outcomes, including brain injury (Synnes et al., 2001), then all unnecessary treatment should be avoided. Deciding when and whether to treat hypotension relies on an understanding of the relationship between BP and brain functioning.

The ability to assess brain activity as a surrogate marker of adequate oxygen delivery may be an important adjunct to decision making in newborns with low BP.

2.2 Measures of physiological and neurodevelopmental health

The existing lack of knowledge in the field of hypotension for preterm neonates has made it very difficult to tackle the problem of low BP. In order to assess whether a chosen treatment approach is beneficial in the long run, it is necessary to use some metric which will allow the health status of the neonate to be characterised. As it can be seen from Figure 2.2, the well-being of the newborn can be quantified by different measures at various stages of treatment. The diagram represents the infant course in the NICU through to the neurodevelopmental follow-up at 2 years of age.

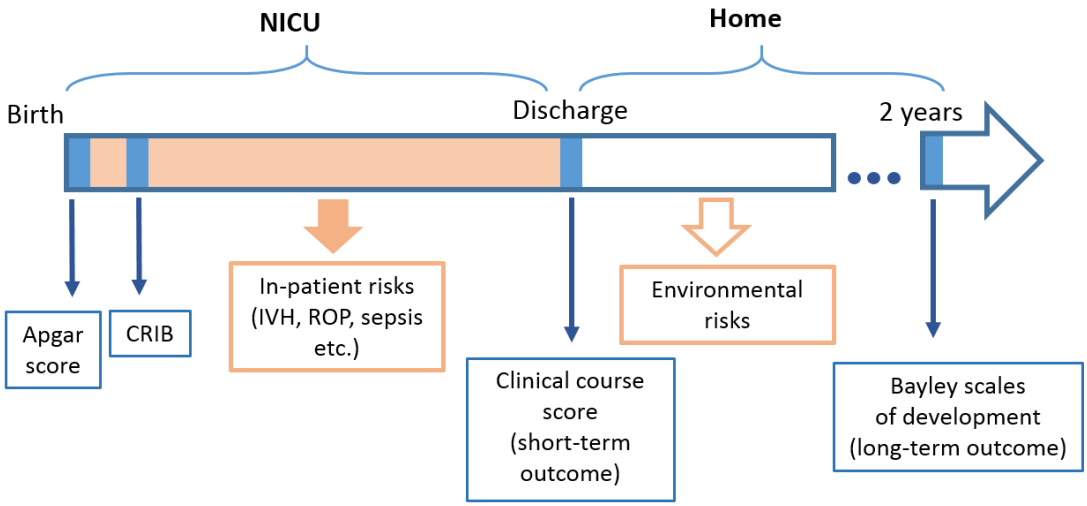


Figure 2.2: The infant's course in the NICU through to the neurodevelopmental follow-up at 2 years of age.

The illness scores are widely used for neonates as they allow for a standardised comparison to be performed between different hospitals or population of patients. There are several different scoring systems which were designed for neonates in order to predict their morbidity and mortality. Frequent properties of neonatal scores include 1) usefulness for a different group of neonates; 2) early applicability after hospitalisation; 3) ability to reproduce prediction of morbidities and mortality and 4) being easy to use (Dorling et al., 2005). Sometimes scoring systems consist of too many variables and may be difficult to complete. In this case, a common practice for investigators is to derive their own health score. This procedure is done by experts using clinical knowledge to select reliable variables and their relative weights if needed. A brief description of the main disease severity scoring system for neonates is provided below.

Apgar score

Apgar score was introduced as the first one to describe newborn health status in the delivery room (Apgar, 1953). This score is based on five factors: respiration, skin colour, muscle tone, reflex irritability and pulse /heart rate and ranges from 0 to 10. It is usually assigned at 1 and 5 minutes after birth and is repeated if the score remains low. Apgar is easy to determine and is routinely used in clinical practice nowadays. At the same time studies (Manley et al., 2010; O'Donnell et al., 2006) identified that despite the consistency of the provided information there is a high interobserver variability of the Apgar score at 5 minutes after birth. Another study argued about the suitability of Apgar for preterm neonates (Manley et al., 2010). This is due to the fact that initial appearance and very early response to resuscitation may be a poor neurological predictor. The Apgar score is also known to be used as an additional variable in other neonatal illness scoring systems (Finn et al., 2016; Richardson et al., 1993).

Clinical risk index for babies (CRIB)

Initially, the performance of neonatal intensive care relied on the risk of mortality adjusted for the birthweight only. CRIB score was developed as an improved tool for assessing neonatal risk in order to compare the performance across different NICUs. This score is comprised of six variables, namely: birthweight; GA in weeks; congenital malformation (excluding lethal malformations); maximum base excess in the first 12 hours; a minimum appropriate fraction of inspired oxygen during first 12 hours; and a maximum appropriate fraction of inspired oxygen during first 12 hours. This score aimed at mortality prediction for neonates less than 32 weeks GA. This score is known to be easy to use and its calculation usually takes five minutes, as compared to other more complex scoring systems (Dorling et al., 2005). This score is calculated during the first 12 hours after birth, which makes it less influenced by the possible treatment procedures.

CRIB II

CRIB II score was developed as an improved version of CRIB (Parry et al., 2003). This score is defined on a scale between 1 and 27 and depends on sex, birth weight, GA, base excess and temperature at admission (Figure 2.3). Higher values are indicative of a greater risk of mortality corresponding to lower GA, birth weight and temperature at admission. Figure 2.3 represents the logistic regression equation for the quantification of the association between CRIB II score and the probability of mortality. CRIB II is intended to improve predictions for very premature neonates and excludes variables which could have been potentially caused by provided care.

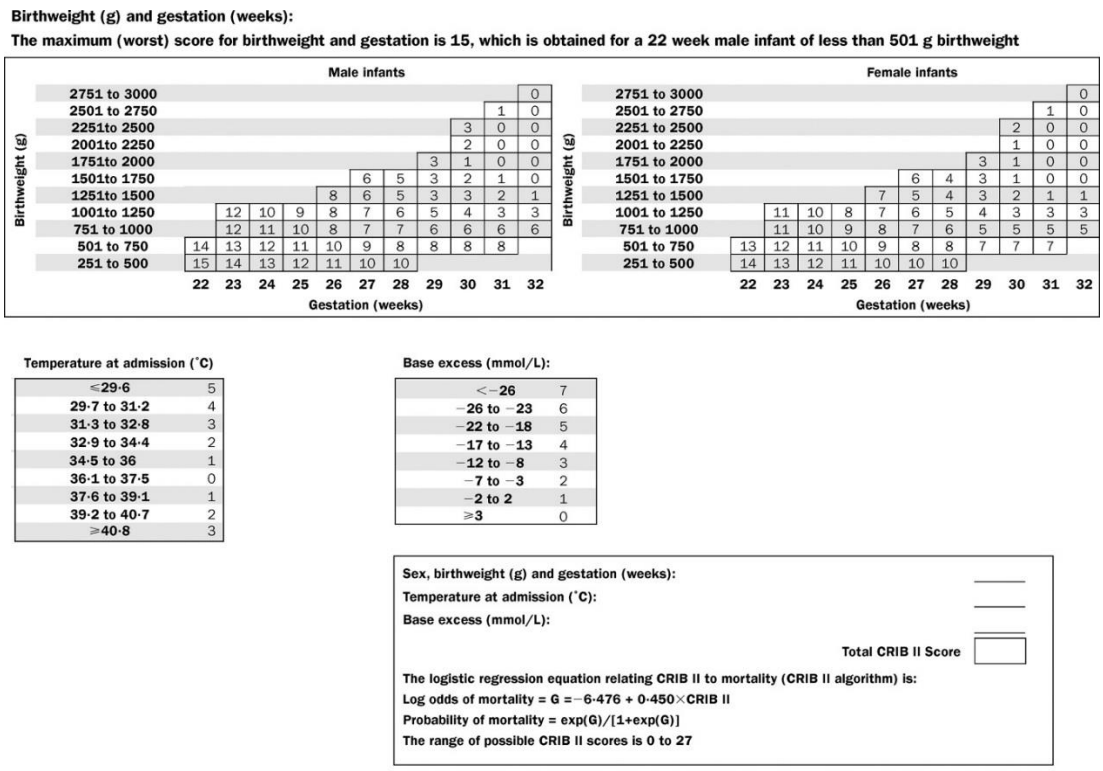


Figure 2.3: Clinical risk index for babies (CRIB II) (Parry et al., 2003).

Clinical course score (CCS)

In this thesis, the short-term outcome of the preterm measured at discharge from the NICU was represented by the clinical course score (CCS). CCS is a score which was designed by the consultant neonatologists (P.F. and E.D.) at Cork University Maternity Hospital. Clinical details and demographics of preterm neonates were collected from the electronic database. Two consultant neonatologists were blinded to both, physiological data and infant identity. The score was then assigned independently to every infant (Lloyd et al., 2016). When grades differed between reviewers, a consensus was reached by discussion. CCS is a binary score, which is based on the discharge summary and medical notes summarising the presence or absence of at least one out of five major neonatal complications: grade III/IV IVH or cystic periventricular leukomalacia; bronchopulmonary dysplasia defined by oxygen dependency at 36 weeks postmenstrual age; necrotizing enterocolitis Bells stage 2b or greater; infection defined as positive blood culture with abnormal inflammatory markers; and retinopathy of prematurity of stage 2 or greater.

Bayley scales of development

The Bayley Scales of Infant Development (Bayley-III), developed by psychologist Nancy Bayley, represent a series of measurements used to assess the development of infants and toddlers aged from 1 to 42 months of corrected age (Bayley, 2006). Bayley-III is an improved

version of the Bayley-II score. It is comprised of five scales with separate composite scores to evaluate cognitive, motor (fine /gross development) and language (expressive /receptive communication) development. Two other scores depend on parental reports. They assess the adaptive and social-emotional behaviour based on the self-control, social responsiveness, communicational and other skills which are estimated on the daily life behaviour. The resulting composite scores are compared to the ones obtained from typically developing children of the same age. Bayley score was initially developed for the Western population whose mother tongue was English. An early neurodevelopmental assessment of preterm neonates is of great importance as it allows for the discovery of possible cognitive, language and motor impairments on early stages. The Bayley scales of infant development were previously used in populations of both term and preterm neonates (Ballot et al., 2017; Greene et al., 2012; Spencer-Smith et al., 2015).

2.3 Heart activity in neonates

The lack of knowledge in the field of hypotension warrants monitoring of other physiological signals including heart activity. This is due to the fact that some severe heart conditions such as low heart rate (bradycardia), heart valve problems and others can lead to low BP and as a result prevent the body to properly circulate the blood. At the same time, HR is also a commonly used parameter for assessing the health status of the neonate.

The normal behaviour of the neonatal heart differs to that of an adult. Recently, many studies have been conducted in order to find the association between heart activity and the health status of the neonate as well as to establish the ranges of normal heart activity (Schwartz et al., 2002). A significant association between the HRV features, the severity of hypoxic ischemic brain injury and the long-term neurodevelopmental outcome at two years of age was reported for 61 full-term neonates (Goulding et al., 2015). Another study (Schwartz et al., 1998) conducted on 33442 newborns found that prolonged QT interval (Figure 2.4) is strongly associated ($p<0.05$) with the sudden infant death syndrome.

For the preterm population, HRV has also shown promising results. The HRV features together with the quantification of general physical movements have been found useful for 2-year outcome prediction in preterms (Dimitrijević et al., 2016). The high frequency component of preterm HRV was shown to be a good biomarker of necrotizing enterocolitis, an acute neonatal inflammatory disease which may lead to death (Doheny et al., 2014). The time-domain HRV measurements have shown a significant difference between septic and non-septic newborns (Bohanon et al., 2015). Lower HRV was observed in children born with low birth weight, including preterm babies (Rakow et al., 2013). HRV has been assessed as a

predictor for successful removal of mechanical ventilation (Kaczmarek et al., 2013); it was demonstrated that babies who failed their first extubation had decreased HRV. A correlation between low frequency oscillations of HRV and BP for preterm neonates has been reported (Rassi et al., 2005). Another study on 92 preterm neonates revealed a significant association between neonatal HRV and respiratory distress syndrome and proposed the use of HRV as an indicator of the morbidity and mortality of the preterm (Cabal et al., 1980). Preterm HRV was also shown to be useful for the prediction of maturation of autonomic control (Jost et al., 2016).

The reported association between heart activity and the health status of the neonate makes HRV a good candidate to provide an insight on the well-being of the newborn during the episodes of hypotension.

2.3.1 Recording heart activity

Electrocardiography (ECG) is a well-established process for recording the electrical activity of the heart. It was shown to be an effective method for measuring heart activity soon after birth (Katheria et al., 2012). This method is non-invasive and routinely recorded in preterms. The ECG detects electrical changes on the skin which are generated by the muscle of the heart by means of electrophysiologic depolarisation and repolarizing during each heartbeat.

A normal ECG is typically characterised by three main components: P wave followed by a QRS complex and a final T wave (Figure 2.4). The P wave represents the depolarization wave, which results in contraction – atrial systole. The QRS complex represents ventricular depolarization. The final T wave corresponds to the repolarisation of the ventricles. Once ventricular depolarisation is completed, the ECG returns to the baseline until the next ventricular repolarisation which produces a T wave.

A lot of changes occur in the normal ECG throughout life. In a similar manner to BP, normal values of many ECG characteristics depend on the age of the patient. More specifically, neonates are known to have a higher heart rate as compared to adults (Schwartz et al., 2002; Selig et al., 2011). It was previously reported that the duration of the QRS complex and P wave increases with the increasing GA of the preterm neonate (Thomaidis et al., 1988). Many of these and other changes are related to differences in physiology, size, and position of the heart and can be indicative of the patient's health status.

Adult ECG is usually recorded using 12 electrodes, which are placed on the limbs and on the chest of the patient. When recording neonatal ECG only three electrodes are used. HRV is usually estimated by calculating the time variation between successive heartbeats. HRV provides a non-invasive assessment of both the sympathetic and parasympathetic control of the heart rate (Akselrod et al., 1981), which are responsible for its acceleration and slowing.

HRV analysis is uniquely suitable for the exploration of the influence of the immature autonomic system on the cardio-respiratory control in preterm neonates. It was reported that HRV, as well as HR, is correlated with GA, where younger preterms were showed to have higher mean HR and lower HRV (Golder et al., 2013).

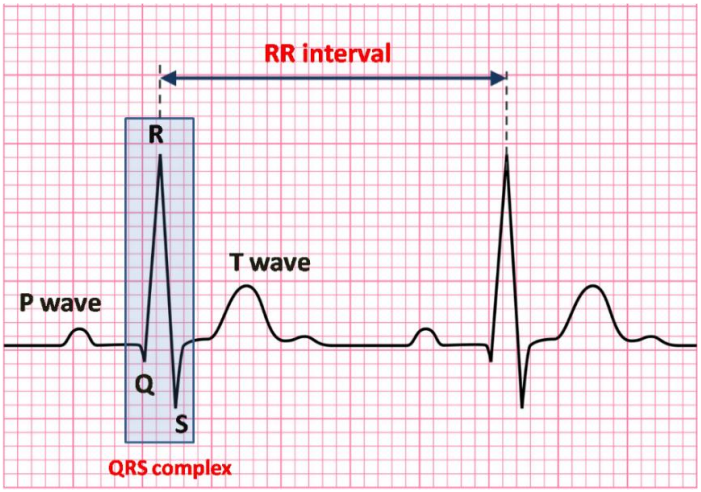


Figure 2.4: A schematic representation of the ECG waveform and its main components: P wave, QRS complex and T wave.

2.3.2 ECG artefacts

In a similar manner to other physiological signals, the ECG can be affected by noise arising from an electrical activity other than a heart origin. The most common ECG artefacts are: 1) respiration artefacts which are introduced by the chest movement during respiration - they are known to affect the low frequencies, particularly 0.4-2 Hz frequency band of the ECG; 2) distortions introduced by the patient movement affect the 1-3 Hz frequency range; 3) muscle tension artefacts affect frequencies in the band from 20 Hz to 150 Hz. The ECG extra-physiological artefacts include electromagnetic inference generated by power lines with the main frequency of 50 Hz, as well as artefacts introduced by the electrodes. The table below summarises the main ECG artefacts and the affected frequency bands.

Table 2.1: Main ECG artefacts with corresponding affected frequency bands.

The source of the noise	Affected frequency band
Respiration	0.4-2 Hz
Movement	1-3 Hz
Muscle tension	20-500 Hz
Power line	50 Hz

2.4 Brain and electroencephalography

As mentioned earlier, the likely way by which hypotension can cause neurological impairment is through reduced perfusion which can lead to insufficient oxygen delivery to the brain tissues.

Unfortunately there is still no direct method to measure CBF and as a result, NIRS and EEG are usually utilized as surrogate measures of brain health, where EEG, in particular, has been shown to be a good predictor of early neonatal outcome (Hallberg et al., 2010).

2.4.1 Brain function and BP

Deciding when and whether to treat hypotension relies on our understanding of the relation between BP, oxygenation and brain activity. The association between MAP and CBF in preterms is controversial. Studies (Jorch and Jorch, 1987; Lou et al., 1979) argued that CBF is pressure passive for preterm neonates of 28-29 weeks GA. At the same time Tyszczuk et al., (1998) found no significant difference in CBF for two groups of preterm infants with MAP above and below 30 mmHg. It was reported that adequate perfusion was maintained at a MAP range of 23.7 to 39.9 mmHg, which is consistent with (Victor et al., 2006b) where EEG activity for MAP values above 23 mmHg was normal.

Several studies have tried to establish the relationship between EEG activity and BP. Some studies have also incorporated measures of cardiac output. D. Shah et al., (2013) identified that BP and EEG energy were associated with flow in the superior vena cava in the first 12 hours of life. However, West et al., (2006) found no association between superior vena cava flow and EEG energy. Increased oxygen extraction has been related to spontaneous activity transients observed in the EEG during the first 6 hours of life (Tataranno et al., 2015). The levels of BP which result in abnormal cerebral activity, as quantified by EEG spectral features and peripheral blood flow measured with NIRS were studied in 35 VLBW infants in (Victor et al., 2006b), where it was reported that a low BP (below 23 mmHg) caused an increase in EEG discontinuity and a decreased relative power of the delta band (0.5-3.5 Hz). Changes in preterm EEG spectral power with maturation were also observed in a study by Niemarkt et al., (2011).

The detection of the presence of a dominant direction for the coupling between BP and EEG can provide a better insight into their mutual interdependency. The neuronal activation of the brain followed by hemodynamic changes has been previously reported in (Mangia et al., 2009). At the same time Caicedo et al., (2016) and Roche-Labarbe et al., (2007) indicated that changes in cerebral oxygenation assessed by NIRS were likely to precede changes in EEG. However, it is worth mentioning that in (Caicedo et al., 2016), the EEG was recorded using only the C3-C4 channel, which does not allow for full coverage of the preterm brain (Pavlidis et al., 2017).

While CBF acquired via NIRS can be used to assess the adequacy of cerebral oxygenation, the evidence that auto-regulation or reactivity of BP and CBF is present in the very preterm baby

is conflicting (Greisen, 2005). Therefore, EEG remains the best tool for the assessment of cortical brain activity in the preterm neonate and may serve as an indicator of cerebral perfusion.

Different physiological modalities have also shown promise in predicting the neurodevelopmental outcome (Lloyd et al., 2016; Périvier et al., 2016). A quantitative analysis of EEG, HR, peripheral oxygen saturation and other clinical features from 43 preterm infants was conducted. Logistic regression of all combined measures showed the potential for the prediction of both mortality and 2-year outcome (Lloyd et al., 2016).

Finding an association between BP and cortical activity with respect to the well-being of the preterm will provide additional knowledge in the field of hypotension. This consequently will help to improve management of low BP in preterm neonates.

2.4.2 Recording EEG

The brain is the most complex organ in the human body, which is responsible for the control of other organs. It consists of two classes of cells, called neurons and glial cells. Glial cells help to maintain homeostasis and also are responsible for the support and protection of the neurons (Kandel et al., 2000). Neurons are considered to be the most important cells in the brain due to their unique ability to send signals to other neurons as well as to the different parts of the body. These cells are interconnected through axons and dendrites, which forms more than 60 trillion neural connections in an adult brain (Stiles and Jernigan, 2010). The main function of the axon is to transmit signals to other neurons by means of a synapse. This allows for cell-to-cell communication.

The electrochemical communication between neurons generates the electric field in the brain tissue. It is very difficult to reliably detect a burst of a single neuron. However, when a large number of neurons perform a synchronised activity, the generated electric field can be captured from the scalp using the technique known as EEG. The EEG recording is obtained by attaching electrodes on the scalp of the patient at specific locations. In order to reduce the impedance caused by dead skin cells, the scalp is prepared and a conductive gel or paste is usually applied on the scalp. Tiny signals detected by the electrodes are then amplified using a differential amplifier.

According to the internationally recognised 10-20 system the electrodes are placed over different areas of the brain (frontal, temporal, parietal, and occipital) and are named accordingly. More specifically, an electrode F2 would be placed over the right frontal part of the brain. Evenly numbered electrodes (2, 4, 6, 8) refer to the placement on the right hemisphere of the brain, whereas odd numbers (1, 3, 5, 7) correspond to the left hemisphere.

When recording EEG in neonates, a reduced number of electrodes is used. This is due to the fact that the skull of a newborn infant is much smaller than that of an adult. As a result, the international 10-20 system of electrode placement adjusted for neonates includes only 9 electrodes: F4, F3, C4, C3, T4, T3, O1, O2, and Cz, as shown in Figure 2.5.

The voltage of the recorded EEG trace is the potential difference between two electrodes, where one of the electrodes is used as a reference electrode. There are different ways to select a pair of EEG electrodes in order to create a channel. Combination of different channels is called montages. A bipolar (or sequential) montage is known to be most commonly used in neonatal EEG monitoring. In a bipolar montage, every channel is represented as a difference between two adjacent electrodes. All channels are created in the form of a chain, where electrodes are linked sequentially. In a referential montage, each electrode is linked to a common reference. In this thesis the analysis was performed on 8 bipolar EEG channels: F4–C4, C4–O2, F3–C3, C3–O1, T4–C4, C4–Cz, Cz–C3 and C3–T3. The bipolar channels for a neonate are also shown in Figure 2.5.

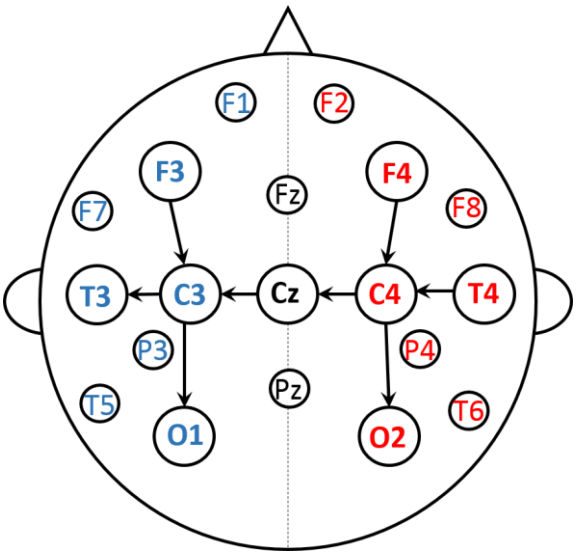


Figure 2.5: The 10-20 international system of the EEG electrodes placement adjusted for the neonates is showed with arrows.

2.4.3 EEG artefacts

The recorded electrical activity which is present in EEG, but which is not of cerebral origin is considered to be an artefact. EEG artefacts are categorised as either physiological or extra-physiological artefacts. Physiological artefacts are generated by sources other than the cerebral cortex and can be either due to movement or bioelectrical potentials. The main EEG artefacts are described below.

Muscle artefact is the most common EEG artefact which is generated by the electrical activity arising from muscles. These artefacts are usually of short duration, high amplitude and with frequency content higher than background EEG activity.

Cardiac artefact is caused by heart electrical activity. This artefact is the most prominent when the neck of the newborn is short. It can be identified by its fixed periodicity and morphology which correlates with the ECG trace. ECG artefacts usually occur on the electrodes which are on the lower part of the head. The cardiac activity can be also expressed as a pulse artefact. This happens when an electrode which is placed over the pulsating vessel is physically moving.

Respiration artefacts are caused by the human respiratory system. If the patient is intubated, which is usually the case for the sick premature neonate, the slow waves are present in the EEG which are correlated with the action of the respirator.

Non-physiological artefacts arise from the environment, more specifically from the nearby devices which generate an electric field. 50 Hz alternating current contamination is one of the most common EEG artefacts. It is easy to identify and can be eliminated using a notch filter. While recording EEG it is possible for the electrode to fully or partially detach from the patient's head. A poor contact of the electrode can result in sharp of slow waves of different morphology and amplitude. The ability of the skin and electrode to function as a capacitor and store an electrical charge across the gel creates a source for the electrode pop (spontaneous discharge) artefact. This artefact has a specific morphology of a very steep rise and a shallow fall with amplitude much greater than EEG activity (White and Cott, 2010). Other possible sources of external signal corruption are wire movement, movement in the environment (near-by ventilator, rocking cot, etc.) and others.

2.4.4 Characteristics of preterm EEG

Lacking proper development, premature babies are especially vulnerable and often are at high risk of complications. The immature brain of the preterm neonate is unique and very different to that of a full-term neonate of an adult. The EEG signal of the neonate contains complex spatiotemporal information which is very difficult to interpret (Pavlidis et al., 2017). As a result, in many NICUs around the world, a simplified version of EEG with a restricted number of channels (1-2) was introduced. This methodology is known as amplitude integrated EEG or aEEG. While providing a good baseline EEG activity (Davis et al., 2015; Wikström et al., 2012), this method does not provide any spatial information, such as specific waveforms and their evolution, which is known to be crucial for the preterm population (Kato et al., 2011; Rennie et al., 2004). The baseline EEG patterns of neonates evolve with maturational changes

taking place within the brain. Therefore, in order to properly interpret neonatal EEG, it is necessary to account for the appropriate maturation features of the preterm neonates with different GAs (Wallois, 2010). The electrical brain activity is usually divided into four frequency sub-bands: delta (0-3.5 Hz), theta (4-7.5 Hz), alpha (8-13 Hz) and beta (13-30 Hz) (Niedermeyer and Silva, 2005).

Discontinuous activity

The activity of the immature brain can be characterised by two interchanging modes of activity: high-voltage activity known as burst or spontaneous activity transients (SATs) followed by longer duration low-voltage activity or inter-bursts intervals (IBI) (Vanhatalo et al., 2005) (Figure 2.6). Unlike the adult population where a similar pattern of burst-suppression is associated with severe brain injury (Lewis et al., 2013), for the preterm infants it is known to be indicative of normal neurological development. The shorter duration of IBI is associated with the increased cortical folding and therefore characterises maturation (increasing GA) (Biagioni et al., 2007). With maturation the overall amount of discontinuity decreases and EEG activity becomes continuous (Niemarkt et al., 2010). According to André et al., (2010) the duration of IBI depends on GA and can be sub-divided into: <60 seconds for 23 – 27 weeks GA; ≤30/40 seconds at 28-29 weeks GA; ≤20 seconds at 30-31 weeks GA; ≤10-15 seconds at 32-34 weeks GA and <10 at 35-36 weeks GA.

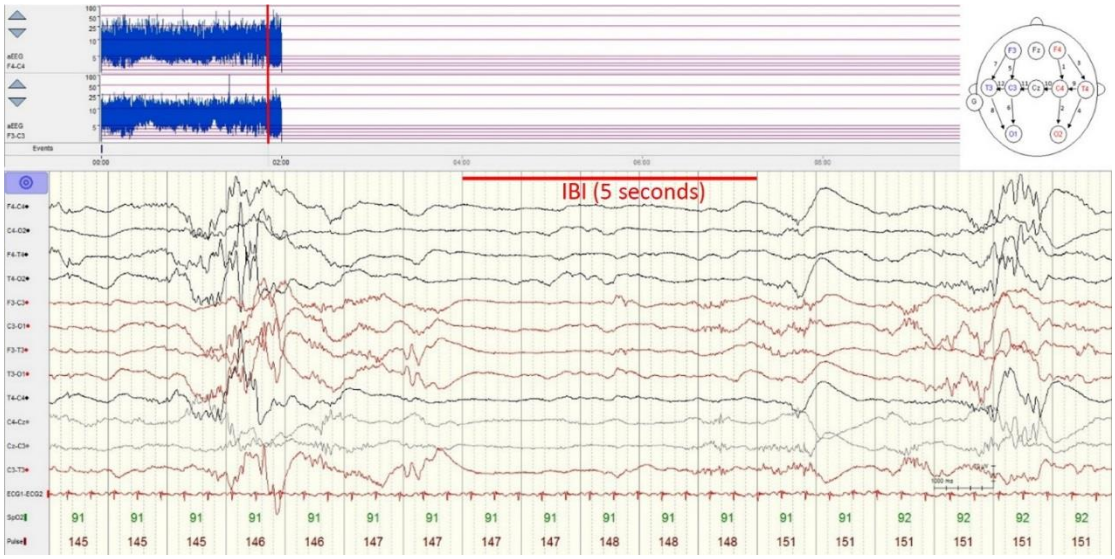


Figure 2.6: Preterm neonate, 31 weeks GA. The figure represents a typical discontinuous activity with high amplitude bursts and low amplitude IBI (Pavlidis et al., 2017).

It was previously reported that before 30 weeks GA infants are in an indeterminate sleep state (André et al., 2010) and proper sleep-wake cycling can be distinguished only after 30 weeks GA (Vecchierini et al., 2007). This information should be taken into account as it allows for the estimation of the health status of the neonate for a specific age.

Delta activity (0-3.5 Hz)

A common background activity for preterms with GA<28 weeks is high amplitude (>300 μV) low frequency activity. Delta activity (0-3.5 Hz) is known as the most important feature of the preterm EEG (Pavlidis et al., 2017) which evolves with maturation and disappears between 38 and 42 weeks GA (André et al., 2010). Several studies have reported a decrease in the amplitude of delta waves with increasing GA (André et al., 2010; Vecchierini et al., 2007).

Theta activity (4-7.5 Hz)

Sharp theta activity on the occipitals of premature infants was shown to be lower in amplitude and faster in frequency as compared to delta waves (Pavlidis et al., 2017). Occipital theta activity is very distinctive in infants with GA <28 weeks. Temporal theta activity (4.5-6 Hz) is the most common at 29-31 weeks GA and disappears at 32-34 weeks GA. Similarly to the delta activity, these changes are also associated with the development of cortical folding.

Synchrony

Synchrony in EEG is present if similar patterns occur simultaneously (with time difference less than 1.5 seconds) in different regions of the brain (Figure 2.7). This feature reflects maturation and interrelation between two hemispheres (Scher, 1996). Although synchrony is known to increase with maturation (André et al., 2010), 88% of IBI were reported to be synchronous between 24 and 27 weeks GA (Vecchierini et al., 2007). Delta wave synchronization was also reported to be present starting from 26-27 weeks GA. Despite the fact that the brain is very immature, infants < 30 weeks GA exhibits “sypersynchrony” (Scher, 1996). Physiological asynchrony emerges after 30 weeks and persists until 36 weeks GA.

Alpha (8 Hz-13 Hz) and beta (13-30 Hz) activity

EGG patter of preterm infants is known to gradually change from high amplitude low frequency waves to low amplitude high frequency waves (Niemarkt et al., 2011; Pavlidis et al., 2017). This leads to the decrease of absolute and relative delta spectral power and increase of relative alpha and beta spectral powers. It is therefore crucial to analyse full spectrum of preterm EEG activity (both low and high frequencies) in order to provide a comprehensive insight into the brain maturational changes and well-being of the preterm.

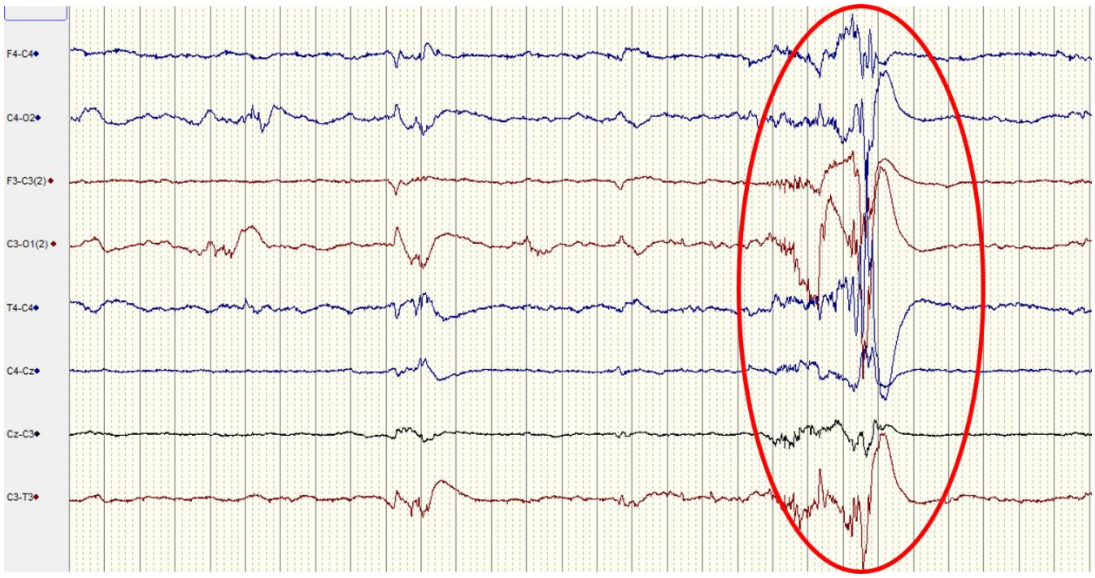


Figure 2.7: Preterm neonate, 26 weeks GA. The figure represents synchrony of high amplitude bursts (highlighted).

2.5 Conclusion

This chapter discusses hypotension in preterm neonates along with the current diagnosis methods, treatment procedures and its implications of the health of the infant. Various clinical metrics which allow for the characteristic of the health status of the neonate were presented. The chapter also provided the main information about other vital physiological signals, namely ECG and EEG, which are commonly used to assess the health status of the neonate. These systems have been previously shown to be affected by hypotension and therefore, may be used in conjunction with BP in order to provide a better insight of the well-being of the newborns during episodes of low BP.

Chapter 3: Methods - biomedical signal processing and machine learning

Deciding when and whether to treat hypotension relies on the understanding of the interrelation between BP and the vital organs such as the heart and brain, as well as their association with the health status of the preterm neonate. This chapter will describe the main signal processing and feature extraction techniques which were used to derive the informative values (features) that characterise the measured data. The chapter will also present the measures which were used for the quantification of linear and nonlinear coupling between physiological signals. In order to decide which ML technique is the most suitable for the problem of prediction of preterm well-being, a brief overview of the current state of the art machine learning algorithms is presented.

3.1 Signal processing and feature extraction

In order to extract informative features that characterise physiological signals of the preterm neonate, it is necessary to take into account the properties of the immature systems and organs which are being developed.

3.1.1 Blood pressure

In clinical practice, the MAP measure is defined as the perfusion pressure of organs in the body and is used most commonly to determine the intervention criteria for the preterm neonates. Both SP and DP are known to be less robust to errors as compared to MAP (Weindling, 1989). In this study, MAP was quantified as $MAP = DP + 1/3\ (SP - DP)$. Figure 3.1 represents traces of SP and DP along with the corresponding MAP.

In this work, the SP and DP signals used for the derivation of MAP were recorded invasively via an umbilical arterial catheter, processed and downsampled in the Philips Intellevue MP70 machine to provide BP recording at 1-second intervals. The MAP is known to be a slowly evolving signal with all significant frequency components within the 0-0.5 Hz band. Low frequency components of MAP (0.005 Hz to 0.16 Hz), which correspond to different components of vasomotion control, were previously investigated for premature infants

(Vesoulis et al., 2017). In the study of Omboni et al., (1993) authors assessed the accuracy of BP recordings performed by finger cuff and compared it with gold-standard inter-arterial BP recordings in adults. The BP was divided into low frequency (0.025 to 0.07 Hz), mid frequency (0.07 to 0.14 Hz) and high frequency (0.14 to 0.35 Hz) components. In another study (Aletti et al., 2013), the effect of respiration on high frequency components of the MAP was investigated, where the high frequency was defined as 0.25 Hz. In all cases, the highest frequency considered are well below 0.5 Hz.

In order to discard artefacts in the MAP signal, MAP values less than 10 mmHg which occur when the pressure transducer is briefly disconnected or moved, were removed. Sharp and non-physiological changes in MAP were automatically eliminated by removing outliers in every 1-hour epoch. To remove the artefacts caused by intervention (e.g. due to infusion given through the line) 10 minutes of MAP before and after each intervention are ignored.

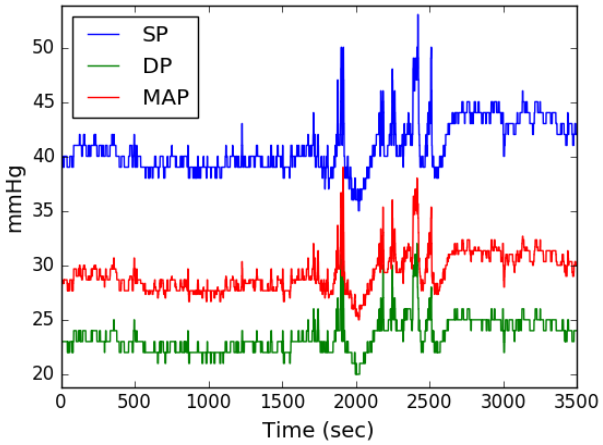


Figure 3.1: Traces of raw systolic (SP) and diastolic (DP) blood pressure along with the corresponding mean arterial pressure (MAP) from the preterm neonate (28 weeks GA).

3.1.2 EEG analysis

While other studies have mainly analysed preselected short EEG epochs, the main strength of this work lies in the analysis performed on long duration unedited multichannel EEG recordings. Quantitative analysis of the EEG data can provide a valuable information about the cerebral activity (Pavlidis et al., 2017). The influence of artefacts was minimised by the following procedures. Prior to preprocessing and feature extraction the usual amplitude-based thresholding of EEG was performed to automatically remove zero-signal and high amplitude artefact (e.g. eye blinking, electrode moved/disconnected). Also, the bipolar montage is used which is known to reduce the effect of some artefacts, e.g. ECG artefacts (Yamada and Meng, 2012).

Prior to EEG feature extraction, the EEG signal is filtered to the range of 0.3-30 Hz and down-sampled to 64 Hz. The EEG is then segmented into 1-minute epochs with the 1-second shift. This segmentation approach allowed to obtain a high resolution insight into the non-stationary EEG signal.

Time domain features

Time domain EEG features included Hjorth parameters - activity, mobility, complexity (Hjorth, 1970) and the number of zero crossings. These features perform simple statistical analysis of the EEG and /or its first and second derivatives. Zero crossing rate was calculated as a rate of sign changes along the EEG. The Hjorth parameters were specifically designed for the EEG analysis. These features were previously used for the grading of HIE in neonatal EEG (Pressler et al., 2001) and therefore may be also indicative of the health status of the preterm infant. Given the signal *X*, the Hjorth parameters are obtained as follows:

$$Activity = {\sigma^2}_X \tag{3.1}$$

$$Mobility = \sigma_{\Delta X}/\sigma_X \tag{3.2}$$

$$Complexity = \frac{\sigma_{\Delta^2 X}/\sigma_{\Delta X}}{Mobility} \tag{3.3}$$

Here ${\sigma^2}_X$ and σ_X are the variance and standard deviation of *X*; ΔX and $\Delta^2 X$ correspond to the first and second derivatives of *X*.

Frequency domain features

Spectral analysis of the EEG is very important as it can provide a better insight into the brain function of the neonate. Studies conducted on both full-term and preterm neonates (Bell et al., 1990; Suppiej et al., 2017) reported that power spectral analysis of the EEG is associated with the neurodevelopmental outcome of the infant. In our study, twelve features were extracted using frequency domain techniques.

Spectral features were estimated on each EEG channel $x_c(t)$, for $c = 1, 2, \dots, 8$. Each epoch of the preprocessed EEG was transformed into the frequency domain using the Discrete Fourier Transform (DFT). The power spectral density $X_c^e(f)$ for the e^{th} epoch of the c^{th} channel was subdivided into four frequency bands: 0.3–3 Hz, 3–8 Hz, 8–15 Hz and 15–30 Hz. This division slightly differs from the standard delta (0.5–3.5 Hz), theta (4–7.5 Hz), alpha (8–12.5 Hz) and beta (13–30 Hz) frequency bands. This was proposed as it better captures the brain dynamics (Vanhatalo et al., 2005), (Tokariev et al., 2012), (Tolonen et al., 2007) and accounts for rapid maturation changes (Pavlidis et al., 2017) in the premature brain. The most important preterm information is known to be concentrated in the lower frequencies of the delta and theta

sub-bands (André et al., 2010), and therefore the gamma (30-100 Hz) band is not used in this study. The power was calculated as total energy within the each sub-band:

$$X_c^e(b) = \int_{f_1(b)}^{f_2(b)} X_c^e(f)df \tag{3.4}$$

where $f_1(b) - f_2(b)$, for $b = 1, 2, 3, 4$ is one of the four frequency bands of the e^{th} epoch in channel c. The relative power was calculated by normalisation of the power in the sub-band by the total power in a given epoch.

The spectral entropy (SE) feature was also extracted for each sub-band. This information theory feature aims to quantify the complexity of the EEG. It was previously reported (Murphy et al., 2015) that SE is significantly correlated with the GA of the preterm neonate and therefore can be indicative of the maturation changes and normal brain development. The probability density function (PDF) P_e was calculated by normalizing the power spectral density of the EEG sub-band for the e^{th} epoch by the total power in the given sub-band. Spectral entropy is calculated as follows:

$$SE(e) = -\sum_{i=1}^{N_f} P_e(i) \log P_e(i) \tag{3.5}$$

where i is a frequency index and N_f is a number of frequency bins in a given sub-band for the epoch. Sixteen EEG features are listed in Table 3.1. For each EEG feature, the median value across all eight channels was calculated in order to reduce the effect of possible focal artefacts.

Table 3.1: EEG features.

Domain	**EEG features**
Time	Activity, mobility, complexity, zero crossing.
Frequency	Relative power (RP), power (P) and spectral entropy (SE) in 0.3–3 Hz, 3–8 Hz, 8–15 Hz and 15–30 Hz bands.

3.1.3 Heart rate variability (HRV) analysis

As discussed in Chapter 2, the heart activity is associated with the outcome of the neonate. Abnormal heart function may alter the blood transportation to the vital organs and systems. In order to extract the HR from the ECG trace, the QRS complexes (Figure 2.4) of the ECG should be identified. This task is usually achieved using an automated QRS detection algorithm.

QRS detection

Most of the energy of the ECG signal is known to be concentrated within the QRS complex (Thakor et al., 1984) in the frequencies between 5 Hz and 20 Hz. As a result, a bandpass filter

with similar cut-off frequencies is usually first applied to the ECG; this substantially reduces the noise (Table 2.1), while preserving the most informative content of the signal. Table 3.2 represents the frequency bands used for the detection of the QRS complex in different algorithms. Most of the QRS detection algorithms can be divided into two stages, namely, preprocessing and the event classification stages. During the preprocessing stage, linear and nonlinear filtering is applied to the ECG signal. In the classification stage, each event is classified as whether or not it is a QRS complex.

Table 3.2: The frequency bands for the detection of the QRS complexes previously proposed in the literature (Elgendi et al., 2010).

Proposed frequency band in literature	**Passband**
(Thakor et al., 1983.) and (Chen and Chen, 2003)	5-15 Hz
(Pan and Tompkins, 1985)	5-12 Hz
(Cuiwei et al., 1995)	8-8.5 Hz
(Sahambi et al., 1997)	3-40 Hz
(Benitez et al., 2000)	8-20 Hz
(Moraes et al., 2002)	9-30 Hz
(Mahmoodabadi et al., 2005)	2-40 Hz

In this thesis, the R peaks of the ECG signal were identified using the Pan-Tompkins method (Pan and Tompkins, 1985), which is considered to be the most accurate in terms of the lowest number of false positives. This algorithm was evaluated on the MIT/BIH database and failed to properly detect only 0.675 percent of the beats. The Pan-Tompkins algorithm can be divided into several stages (Figure 3.2).

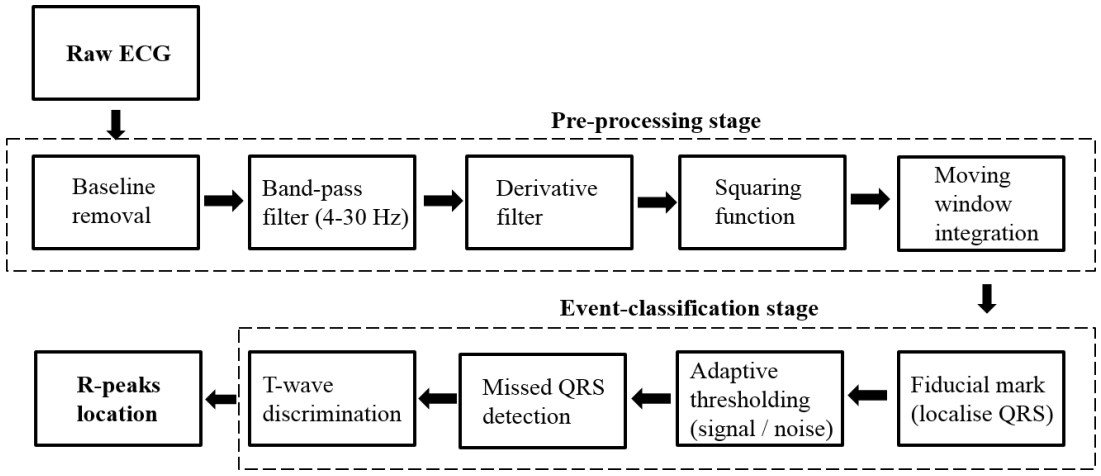


Figure 3.2: Stages of the Pan-Tompkins algorithm.

1. Pre-processing stage

In the first stage, the ECG signal is bandpass filtered to attenuate muscle noise, 50 Hz interference, T wave inference as well as baseline wander. The normal range of the HR changes with maturation. More specifically, a normal HR of a neonate ranges from 90 to 165 beats per minute (bpm), whereas the resting heart rate of an adult is 55 – 90 bpm (Fleming et

al., 2011). In order to account for this difference and modify the algorithm accordingly, the ECG signal was bandpass filtered using 4-30 Hz cut-off frequencies instead of the originally proposed 5-15 Hz. Increasing the low-pass cut-off frequency to 30 Hz allows to adjust the algorithm for neonates by emphasising the R peaks in order to better distinguish between the R peak and the P wave. The output of this filter can be seen in Figure 3.3 (Filtered ECG stage). The comparison of the QRS detector output using the originally proposed 5-15 Hz cut-off frequencies is represented in Figure 3.4. It can be observed that after bandpass filtering, the R peaks are merged with the T and P waves and therefore cannot be properly identified.

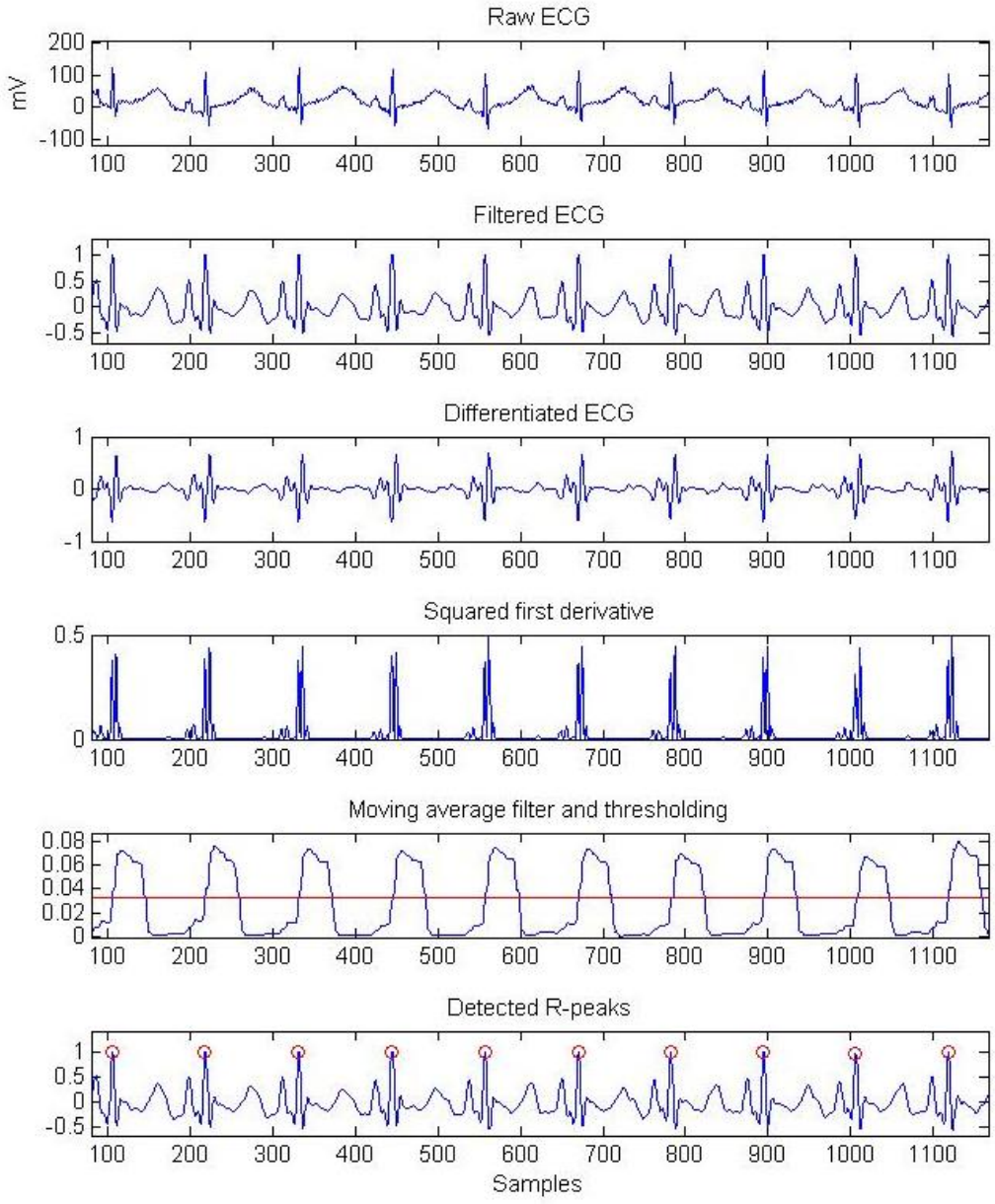


Figure 3.3: An example of the preterm ECG signal going through different stages of the Pan-Tompkins R peak detection. Bandpass cut-off frequencies are 4-30 Hz.

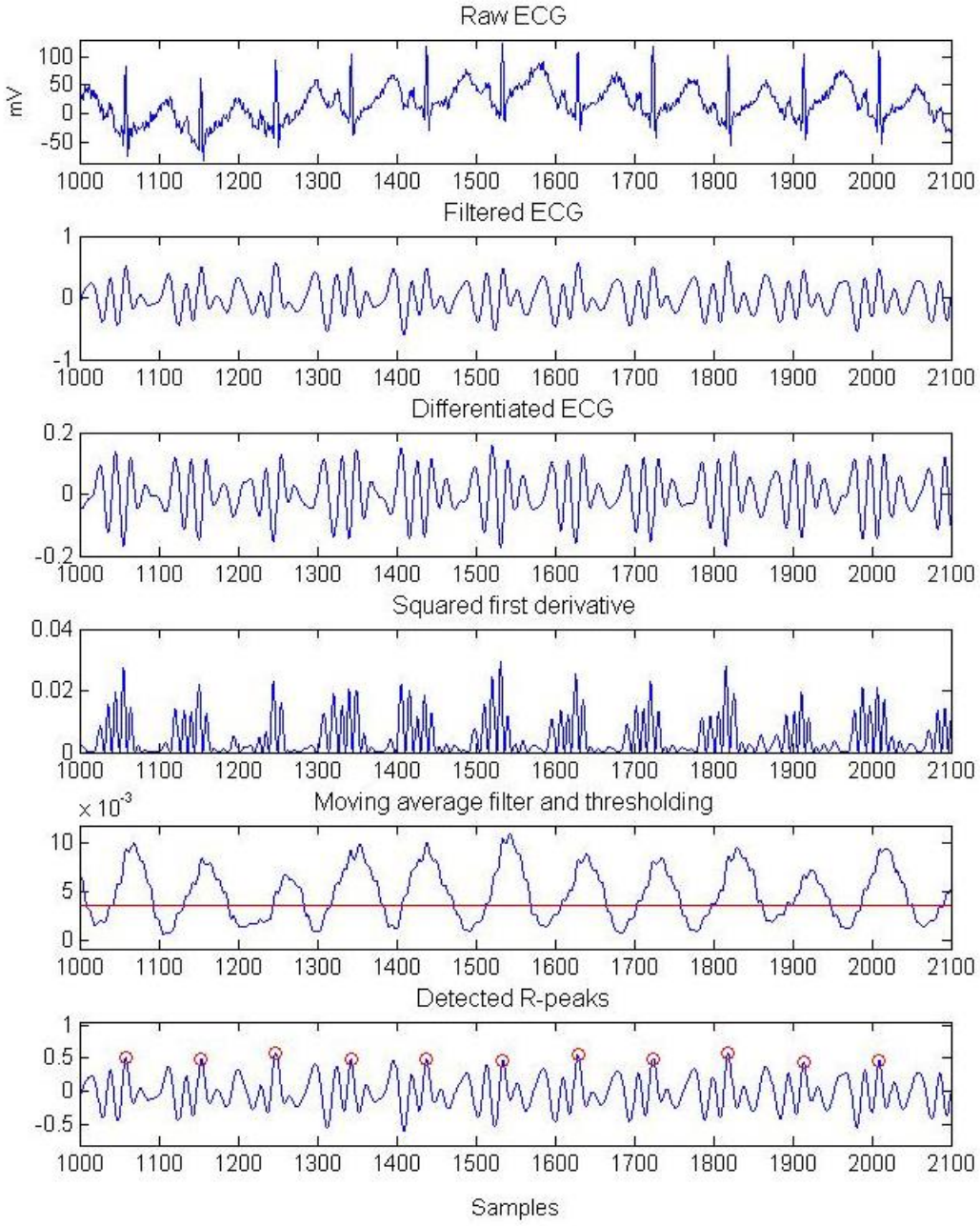


Figure 3.4: An example of preterm ECG signal going through different stages of the Pan-Tompkins R peak detection. Bandpass cut-off frequencies are originally proposed 5-15 Hz.

In order to enhance the slope of the QRS complexes the filtered ECG signal is then differentiated. This is followed by squaring the signal, point by point, in order to nonlinearly amplify the output of the differentiation step and emphasize higher frequencies. The slope of the R wave does not provide sufficient information for the isolation of the QRS complex. In order to extract the information about the width of the QRS complex, a moving average filter is applied. Generally, the moving average window should be approximately the same as the widest QRS complex. A window which is too wide can cause the QRS and T complexes to merge together. Similarly to (Pan and Tompkins, 1985), the width of the moving window was

set to 150 ms, which is an average duration of the QRS complex and is sufficiently small to avoid possible contribution of the T wave.

2. Event classification

During the event classification stage, the algorithm produces a pulse-shaped waveform. The classification of the QRS complex can be divided into several stages: 1) Fiducial mark is used for the temporal location of the QRS complex. The R peak is determined from the rising edge of the obtained pulse waveform according to the maximal slope. 2) In order to determine whether this pulse corresponds to the R peak, a highly-sloped T wave or a noise artefact, adaptive thresholding is performed. The algorithm uses two threshold values to classify each non-zero sample as either noise or signal. Thresholds are constantly updated based on the new estimations of the signal and noise thresholds. 3) In order to reduce the number of false negatives, the algorithm performs a back search in order to find missed QRS complexes, which allows for adaptation to an irregular heart rate. This step is activated if no QRS complexes have been detected for longer than 1.66 times the average RR interval (Pan and Tompkins, 1985). 4) False positives are reduced by excluding all peaks which precede or follow larger peaks by less than 200 ms (this corresponds to 300 bpm). 5) In order to discriminate a T wave, a check is performed for the RR intervals which are greater than 200 ms and less than 360 ms. This is implemented by measuring the slope the QRS candidate and comparing it to the slope of the previously detected QRS complex. The output of the QRS detector generates the final RR intervals, which are referred to as the tachogram.

In this study, the HR behaviour was quantified using the time-domain and frequency-domain analysis. Abnormal values of time intervals between R peaks (RR intervals) caused by artefacts were corrected by the moving average filter or discarded if the epoch was too corrupted. The corrected RR intervals, NN intervals, were used to estimate the instantaneous heart rate signal. Thirteen HRV characteristics were extracted from the RR intervals. These features have been previously used for various purposes to quantify the HRV in term and preterm infants (Temko et al., 2015), (Selig et al., 2011).

3.1.4 Time domain HRV features

The time domain features were derived using the simple statistics of the RR interval distribution and include the mean of the RR interval (MeanRR); skewness and kurtosis of the RR distribution; the standard deviation of the RR interval (SDNN), which reflects all the periodic components responsible for the variability in given epoch and is calculated as follows:

$$SDNN = \sqrt{\frac{\sum_{i=1}^{n}(NN_i - \overline{NN})^2}{n-1}} \tag{3.6}$$

where NN_i is the time interval between i^{th} to the $i+1$ R peaks, $\overline{NN}$ is the average interval given n intervals in total. Decreased SDNN was previously associated with an increased grade of HIE in full-term neonates assessed during the first 48 hours after birth (Goulding et al., 2015). For the adult population, this measure was used for the prediction of both morbidity and mortality (Shaffer and Ginsberg, 2017). Decreased HRV was previously associated with hypoxic brain injury in newborns (Matić et al., 2013) and with the failure of the first extubation in the preterm infants (Kaczmarek et al., 2013).

Triangular interpolation of the RR histogram (TINN) is a geometric method to represent another time domain feature. Figure 3.5 schematically represents the computation of the TINN using a sample density distribution. This measure is defined as the width of the baseline of an imaginary triangle that best approximates the histogram of the RR signal using the least squares method (Electrophysiology, 1996). Reduced TINN values were previously reported for encephalopathic neonates. This feature was also demonstrated to be predictive of the neurodevelopmental outcome at 2 years of age (Goulding et al., 2015).

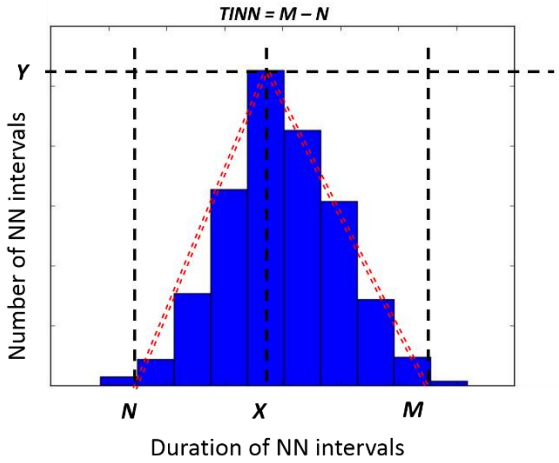


Figure 3.5: Schematic representation of the TINN feature computation. *X* is the bin with the maximum number of NN intervals, *N* and *M* are determined by finding the interpolated triangle that best fits to the histogram.

Other HRV features include the temporal information, such as the root of the mean-squared difference of the successive RR intervals (RMSSD) and the ratio between SDNN and RMSSD. RMSSD reflects the short-term variability as is calculated as follows:

$$RMSSD = \sqrt{\frac{\sum_{i=1}^{n-1}(NN_i - NN_{i+1})^2}{n-1}} \tag{3.7}$$

where NN_i is the i^{th} RR interval, NN_{i+1} is the next RR interval, and n is a total number of RR intervals. RMSSD estimates the short-term component of HRV and represents the parasympathetic activity of the heart. In (Dimitrijević et al., 2016) RMSSD was shown to

improve the 2-year outcome prediction for preterm neonates, with lower RMSSD values corresponding to minor neurologic dysfunction and cerebral palsy. Other studies have reported RMSSD as a good early predictor of a septic shock for adults (Chen and Kuo, 2007) as well as sudden unexplained death in epilepsy (DeGiorgio et al., 2010).

3.1.5 Frequency domain HRV parameters

For the frequency domain method, a power spectral density (PSD) estimate is calculated for the RR intervals. The PSD estimator requires equidistant sampling, therefore, the RR intervals were uniformly resampled at 256 Hz. The PSD was then obtained using the DFT transform. This was followed with HRV quantification by power in various bands: power in very low frequency band (VLF , 0.008 – 0.04 Hz), power in low frequency band (LF, 0.04 – 0.2 Hz); power in high frequency band (HF, 0.2 – 1 Hz); and the LF/HF ratio. The frequency domain features extracted from the PSD were calculated as absolute powers of VLF, LF and HF bands by integrating the spectrum over the limits of the given frequency band.

All frequency domain features used in this thesis were extensively utilised for HRV analysis and are indicative of the physiological status of the patient. More specifically, VLF characterises very slow HR fluctuations which were previously reported as an independent risk predictor for patients with congestive heart failure (Hadase et al., 2004). Some studies reported that VLF may reflect the thermoregulation to ambient temperature changes in adults (Kinugasa and Hirayanagi, 1999) and preterm neonates (Stéphan-Blanchard et al., 2013). Neonatal HRV evaluated by spectral analysis is usually characterised by the dominant activity in the LF band. The LF is mediated predominantly by the sympathetic component. It is also considered to represent the Mayer waves of BP changes (Draghici and Taylor, 2016). Mayer waves are cyclic changes in arterial BP at the frequency of about 0.1 Hz resulting from the oscillation of sympathetic vasomotor tone (Pagani et al., 1996). Reported evidences also suggest that Mayer waves are dependent upon intact arterial baroreflex function, which is the most powerful regulator of BP. Both human and animal studies have supported the finding that a reduced spectral power in the LF component is indicative of the impaired function of the autonomic nervous system (Piccirillo et al., 2009; A. J. Shah et al., 2013). Reduced LF and HF features were reported for neonates with hypoxic ischemic encephalopathy (HIE), implying the reduction of autonomic function (Goulding et al., 2015). Calculation of the LF/HF ratio is a method to establish the ratio between the components of the autonomic nervous system – sympathetic/parasympathetic balance (Electrophysiology, 1996).

3.1.6 Nonlinear HRV analysis

The heart is known to have a complex control system. This allows us to assume that the HRV is generated by nonlinear mechanisms and that nonlinear methods of HRV analysis may provide additional information about the autonomic control of the HR.

Nonlinear HRV features are represented here by the approximate entropy (ApEn) which quantifies the regularity and complexity of stationary signal (Pincus, 1991). Large values of ApEn are indicative of low predictability of fluctuations in the successive RR intervals. Given the sequence $x(n) \in \{x(1), x(2), \dots, x(N)\}$, where N is a total number of data points, the algorithm for ApEn calculation is summarised as follows:

1. Create a set of $N - m + 1$ vectors of m components:

$$X_m(i) = [x(i), x(i+1), \dots X(i+m-1)]$$

Here m is the embedding dimension and vector $X_m(i)$ represents the sequence of m consecutive RR values starting at beat i.

2. Define the distance $d[X_m(i),\ X_m(j)]$ between two vectors as the maximum absolute difference between the corresponding elements:

$$d[X_m(i),\ X_m(j)] = max_{k=0,\dots,m-1}[|x(i+k) - x(j+k)|]$$

3. Defining $n_r^m(i)$ as the number of the $N - m + 1$ vectors $X_m(j)$ which are similar to $X_m(i)$, then

$C_r^m(i) = n_r^m(i)/(N - m + 1)$, where $n_r^m(i) = no.of\ d[X_m(i),\ X_m(j)] \leq r$. $C_r^m(i)$ is the probability to find a sequence of m beats, that is similar to the sequence $X_m(i)$.

4. Take the natural logarithm of each $C_r^m(i)$ and average it over i as defined in the previous step 3:

$$\Phi^m(r) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} \ln C_r^m(i)$$

5. Calculate ApEn for a finite number of points N as follows:

$$ApEn(m, r, N) = \Phi^m(r) - \Phi^{m+1}(r) \tag{3.8}$$

A high degree of regularity represented by low ApEn values implies that the two sequences $(X_m(i),\ X_m(j))$ that are close (similar) within the distance r in an m-dimensional space, remain close for the $(m+1)$-dimensional space. In order to calculate ApEn, it is necessary to choose two parameters, namely, parameter m – the embedding dimension and parameter r – a threshold tolerance. Based on the analysis of deterministic and stochastic processes (Pincus et al., 1991; Pincus and Keefe, 1992) it was suggested to select $m = 2$ and

r in the range between 0.1 and 0.25 times the standard deviation of the given time series. Values $m = 2$ and $r = 0.2 * SD$ are routinely used in most HRV studies and were applied for the ApEn computation in this thesis as well.

3.1.7 Allan analysis

The Allan Factor (AF) is a scale-dependent measure which quantifies the variability of successive counts (Lowen and Teich, 1996). AF in a predetermined time interval T is defined as the ratio of the variance of successive counts divided by twice the mean of the counts.

$$A_i(T) = \frac{\langle [N_{j+1}(T) - N_j(T)]^2 \rangle}{2\langle N_{j+1}(T) \rangle} \tag{3.9}$$

Here i is the epoch number, $N_j(T)$ is the number of RR intervals occurring in a window of length T; j - is the index of the window of length T within the i^{th} epoch; $\langle . \rangle$ is the mean operator. For a random process in which fluctuations are not correlated $A(T) = 1$. When the process is periodic, the variance decreases and $A(T)$ approaches to zero. The application of the AF to HRV data revealed the importance of this measure, more specifically, the AF has been reported to be indicative of whether a patient suffers from the heart failure (Turcott and Teich, 1996). In this study, the AF is calculated for every 5-minute epoch of RR intervals with a time scale of T = 60 seconds. A concise list of the thirteen HRV features is presented in Table 3.3.

Table 3.3: Frequency- and time-domain features extracted from the ECG.

Domain	**HR features**
Time	MeanRR, SDNN, skewness, kurtosis, TINN, RMSSD, SDNN/RMSSD, ApEn, Allan Factor
Frequency	Power in VLF (0.008 – 0.04 Hz); LF (0.04 – 0.2 Hz) and HF (0.2 - 1 Hz) bands; ratio LF/HF

3.2 Modelling interaction between physiological signals

In this study, both linear and nonlinear measures of the interaction of EEG and MAP features were calculated over a 30-minute moving window with a 30-second shift. This window length allows one to focus on the short-term dynamics of both the EEG and MAP signals.

3.2.1 Linear measures of interaction: correlation and coherence

Correlation is the most common way to determine the extent to which two variables co-vary linearly and it is defined as:

$$r_{xy} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}, \tag{3.10}$$

where $\bar{x}$ and $\bar{y}$ are mean values of x and y respectively. This measure captures information on the time coupling and waveform similarity between two signals. Correlation is sensitive to polarity and its values range from -1 to 1, where -1 represents a perfect negative dependence, 1 corresponds to a perfect positive dependence, and 0 implies no dependence between the two variables. Correlation is a powerful tool for measuring linear dependence, however, this measure also has a number of drawbacks. Correlation is not able to capture the nonlinear association between variables, neither can it to account for a phase shift between two signals, which may result in misleading findings. This drawback can be taken into account by introducing a delay τ into one of the sequences. This method is usually referred to as cross-correlation and it measures the dependency between two variables as a function of the displacement of one variable relative to another. The cross-correlation function is defined as follows:

$$C(\tau) = corr\big(x(t), y(t-\tau)\big), \tag{3.11}$$

where $corr$ is a correlation, τ represents the delay between two sequences.

Unlike correlation which measures interaction in the time domain, the coherence measure has an advantage of showing the similarity of two variables for a chosen frequency. At each frequency f the coherence function Coh_{xy} is defined by:

$$Coh_{xy}(f) = \frac{|S_{xy}(f)|^2}{S_{xx}(f)S_{yy}(f)}, \tag{3.12}$$

where $S_{xy}(f)$ is the cross power spectral density between x and y at frequency f; $S_{xx}(f)$ and $S_{yy}(f)$ are the auto-spectral density of x and y, respectively. Calculation of cross spectral density allows coherence to account for the possible lag between signals, whereas correlation is very sensitive to phase lag. The calculation of coherence also involves squaring the signal, thus producing values in the range between 0 and 1.

The Welch method (Welch, 1967) is usually used to estimate the PSD. It is done by dividing the signal into overlapping segments, computing a periodogram for each segment and averaging the result. This method assumes that a signal is stationary, which implies that the statistical properties of the signal do not change over time. The resulting estimate of the PSD of the signal $x(t)$ is computed as follows:

$$P_x(\omega) = \frac{1}{p}\sum_{i=1}^{p} P_{x_i}(\omega) \tag{3.13}$$

where p is a number of overlapping segments x_i, $i = 1,2,\dots,p$; $P_{x_i}(\omega) = |X_i(\omega)|^2$ is the power spectrum of the i^{th} segment. The averaging of the PSD allows for a decrease in the

variance and provide a smoother estimate as compared to a single periodogram estimate performed on all the data.

Both correlation and coherence measure only the linear association between signals. Due to the nature of non-stationary physiological processes, the conditions of data stationarity within a given segment are not always satisfied. As a result, this may lead to a bias in the estimation of the spectral density. Therefore, the limitations of both methods should be acknowledged when working with real physiological data.

3.2.2 Nonlinear measure of interaction: mutual information

Due to the likely complex relation between physiological signals, such as brain function (EEG) and MAP, (Pikovsky et al., 2003), nonlinear methods of interaction were included as well. Mutual information (MI) is an information theoretic measure of dependency between two random variables defined as:

$$MI(X,Y) = H(X) + H(Y) - H(X,Y), \tag{3.14}$$

where $H(X), H(Y)$ are the Shannon entropies for sequences X and Y each labelled into r and c labels respectively. After substitution of:

$$H(X) = -\sum_{i=1}^{r} p(i) log_2 p(i) \tag{3.15}$$

the MI is obtained as follows:

$$MI(X,Y) = \sum_{i=1}^{r}\sum_{j=1}^{c} p(i,j)\, log_2\left(\frac{p(i,j)}{p(i)p(j)}\right), \tag{3.16}$$

where $p(i)$, $p(j)$ are the probabilities of the occurrence of a particular label, i,j, in the sequences X, Y; $p(i,j) = p(i)p(j|i)$, where $p(j|i)$ is the probability that a label, j, occurs in sequence Y, given another label, i, occurred in sequence X. It is easy to show that if the sequences X and Y are independent, then $p(j|i) = p(j)$, and the ratio term $p(i,j)/(p(i)p(j))$ becomes 1 and therefore $MI(X,Y)$ becomes 0. MI is a symmetric measure ($MI(X,Y) = MI(Y,X)$) and unlike correlation or coherence, it quantifies both linear and nonlinear dependences. In a similar manner to other information measures, the most common way to calculate MI from the empirical data is to use histogram binning (labelling) in order to estimate the probability density distribution (Figure 3.6).

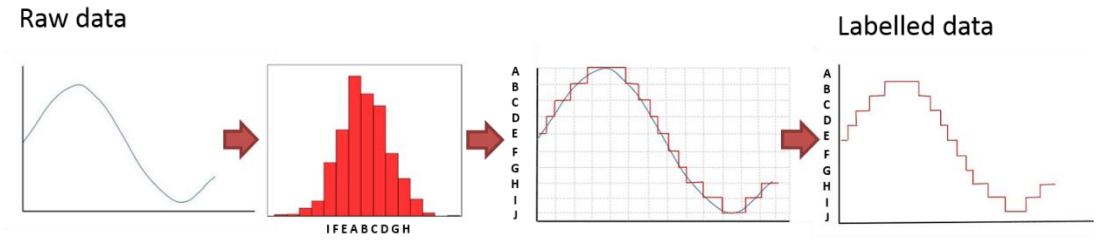


Figure 3.6: Schematic representation of the signal labelling using histogram binning.

The choice of the number of bins into which the two sequences (X, Y) are subdivided is important and may significantly affect the results. If there are too few, then it might be impossible to distinguish any structure in the distribution. Too many bins might result in occupation numbers of zero which provides no meaningful information.

Adjusted mutual information

In order to minimise the effect of the choice of the number of labels, the MI can be calculated as adjusted mutual information (AMI), which unlike conventional MI corrects the effect of the agreement between two sequences which happens solely due to chance (Vinh et al., 2010). In particular, AMI accounts for the fact that MI tends to increase as the number of different labels increases, regardless of the actual amount of interaction between the two sequences. The main advantage of AMI is that for the chosen number of bins, AMI measures the interaction that is adjusted for a random chance; the conventional MI measure increases with the increase of random interactions which are caused by the high number of bins. The AMI for two sequences (X, Y) is computed as:

$$AMI(X,Y) = \frac{MI(X,Y) - E[MI(X,Y)]}{mean[H(X),H(Y)] - E[MI(X,Y)]} \tag{3.17}$$

This method of adjusting a given measure for a random chance using its expected value was previously proposed in (Hubert and Arabie, 1985). The expected values are used to remove the baseline component of the measure. The $mean$ term in the denominator acts as a normalization.

The overlap between sequences X and Y each consisting of N datapoints can be represented in matrix form by the $r \times c$ contingency table $\mathcal{M} = [n_{ij}]_{j=1\dots c}^{i=1\dots r}$, where r and c defines a number of unique labels in X and Y respectively. Here, $\sum_{i=1}^{r} a_i = \sum_{j=1}^{c} b_j = N$; n_{ij} represents the number of common objects in both a_i and b_j. A contingency table $\mathcal{M}$ is shown in Figure 3.7, with $a_i = \sum_{j=1}^{c} n_{ij}$ as a row marginal and $b_j = \sum_{i=1}^{r} n_{ij}$ as the column marginal.

		b_1	$\cdots$	b_j	$\cdots$	b_c
	a_1	n_{11}	$\cdots$	$\cdot$	$\cdots$	n_{1c}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
X	a_i	$\cdot$		n_{ij}		$\cdot$
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	a_r	n_{r1}	$\cdots$	$\cdot$	$\cdots$	n_{rc}

Y

Figure 3.7: A contingency table $\mathcal{M}$ with $r \times c$ size, with a_i and b_j as the row and columns marginal.

Therefore, using the contingency table $\mathcal{M}$, MI is defined as:

$$MI(X,Y) = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{n_{ij}}{N}log_2\frac{n_{ij}N}{a_i b_j} \tag{3.18}$$

The expected value of MI is obtained by summation over all possible contingency tables $\mathcal{M}$ obtained by permutations (Vinh et al., 2009) and is defined as:

$$E[MI(X,Y)] = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{n_{ij}=max(a_i+b_j-N,0)}^{min(a_i,b_j)}\frac{n_{ij}}{N}log_2\left(\frac{N\times n_{ij}}{a_i b_j}\right)\times \frac{a_i!\,b_j!\,(N-a_j)!\,(N-b_j)!}{N!\,n_{ij}!\,(a_i-n_{ij})!\,(b_i-n_{ij})!\,(N-a_i-b_j+n_{ij})!} \tag{3.19}$$

The resulting measure of AMI is normalised, which facilitates interpretation and allows for comparison. More specifically, the values of AMI are close to 0 (small negative values can also occur) when MI is equal to the expected value obtained by chance, as well as for independent X and Y sequences. AMI is equal to 1 when X and Y are identical (i.e. perfectly matched).

3.2.3 Directionality of interaction: transfer entropy

MI does not contain any directional information as it is symmetric measure, where $MI(X,Y) = MI(Y,X)$, and therefore it is not effective at predicting future events from the data, or deriving the causality between two sequences. Transfer entropy (TE) is an extension of MI which takes into account the direction of the informational flow, under the assumption that the underlying processes can be described by a Markov model (Ragwitz and Kantz, 2002). TE allows the quantification of the exchange of information between two sequences, for each direction, by means of an introduced time lag in either one of the sequences. TE from a sequence X to another sequence Y is the amount of uncertainty reduced in future values of Y by knowing the past values of X, given past values of Y. The amount of information transferred from sequence X to sequence Y is denoted as $TE_{(\boldsymbol{X}\rightarrow\boldsymbol{Y})}$ and is computed as follows:

$$TE_{(\boldsymbol{X}\rightarrow\boldsymbol{Y})} = \sum_{y_{t+u}, y_t^{d_y}, x_t^{d_x}} p(y_{t+u}, y_t^{d_y}, x_t^{d_x}) \log\left(\frac{p\left(y_{t+u}|y_t^{d_y}, x_t^{d_x}\right)}{p\left(y_{t+u}|y_t^{d_y}\right)}\right) \tag{3.20}$$

Here $p\left(y_{t+u}|y_t^{d_y}, x_t^{d_x}\right) = p(y_{t+u}, y_t, x_t)/p(y_t, x_t)$ and $p\left(y_{t+u}|y_t^{d_y}\right) = p(y_{t+u}, y_t)/p(y_t)$; t is the point in time and u indicates the prediction time, e.g. y_{t+u} is the value of Y at time $t+u$. Values $y_t^{d_y}$ and $x_t^{d_x}$ are d_y - and d_x - dimensional delay vectors defined as: $x_t^d = (x(t), x(t-\tau), x(t-2\tau), \ldots, x(t-(d-1)\tau))$, with τ as embedding delay. If the two processes are mutually independent there will be no transfer of information, therefore $p(y_{t+u}|\, y_t) = p(y_{t+u}|\, y_t, x_t)$ and $TE_{(\boldsymbol{Y}\rightarrow\boldsymbol{X})} = TE_{(\boldsymbol{X}\rightarrow\boldsymbol{Y})} = 0$.

Binning of the data for the TE computation is no longer sensible as it ignores the neighbourhood relations in the continuous data and destroys the information about the absolute values of the original data (Wollstadt et al., 2017) which is crucial for TE calculation. Examples, where TE estimation fails due to the use of binned time series were previously reported in (Pompe and Runge, 2011). In (Wibral et al., 2013) the estimation of TE was only properly obtained when using continuous data as opposed to its binned version. This problem has been solved by using the Kraskov-Stogbauer-Grassberger (KSG) nearest-neighbor based TE estimator for continuous data (Kraskov et al., 2004). KSG is an improved box kernel estimator, which uses dynamically altered kernel width r which depends on the number of nearest neighbors. TE can be written using the representation of four Shannon entropies as:

$$TE_{(\boldsymbol{X}\rightarrow\boldsymbol{Y})} = H\left(y_t^{d_y}, x_t^{d_x}\right) - H\left(y_{t+u}, y_t^{d_y}, x_t^{d_x}\right) + H\left(y_{t+u}, y_t^{d_y}\right) - H\left(y_t^{d_y}\right) \tag{3.21}$$

Every Shannon entropy $\left(H(.)\right)$ is then estimated by the nearest-neighbor technique and the KSG estimator. The nearest-neighbor technique uses the statistics of the distance between neighboring data points in the embedding space. The KSG estimator uses a fixed number of neighbors K for the search in the highest dimensional space and then projects the resulting distances to the lower dimensional space as the range to look for neighbours (Kraskov et al., 2004). After adapting this technique to the formula with Shannon entropies (Gómez-Herrero et al., 2015), TE can be rewritten as:

$$\begin{aligned} TE_{(\boldsymbol{X}\rightarrow\boldsymbol{Y})} = \psi(K) + \langle \psi\left(n_{y_t{}^{d_y}} + 1\right) - \psi\left(n_{y_{t+u}y_t{}^{d_y}} + 1\right) \\ - \psi(n_{y_t}^{d_y} n_{x_t}^{d_x})\rangle\, t, \end{aligned} \tag{3.22}$$

where ψ is a digamma function and $\langle . \rangle t$ indicates an averaging over different time points and K is a number of nearest neighbours in the highest dimensional space, n number of points within search radius of the lower dimensional space.

3.3 AUC as a measure of statistical predictive power and classification metric

Throughout this work the receiver operating characteristic (ROC) is used for a number of purposes, namely, to assess the performance of the classifier and to quantify the discriminative (predictive) power of features with respect to the health status of the preterm. For a binary classification problem as considered in this thesis, the predicted outcomes are labelled either positive or negative. Consequently, there are four possible outcomes from a classifier (Figure 3.8): true positive (TP) – if the outcome of positive prediction and the actual value is positive; false positive (FP) – if the actual value is negative but the predicted outcome is positive; true negative (TN) occurs when both the prediction and the outcome are negative; and false negative (FN) for the cases when the prediction is negative, while the actual value is positive.

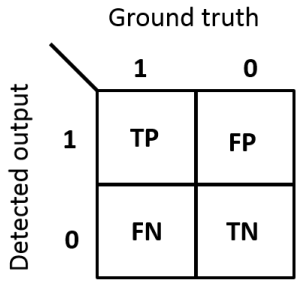


Figure 3.8: Confusion matrix.

The receiver operating curve is constructed with a set of sensitivity and specificity values which are obtained by adjusting the decision-making threshold. Sensitivity refers to the ability to correctly detect a sick patient, while specificity characterises the ability to correctly reject a healthy patient. These characteristics are computed as follows:

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.23}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3.24}$$

The ROC graph plots all sensitivity and specificity pairs resulting from continuously varying the decision threshold over the entire range of results as shown in Figure 3.9. The area under the ROC curve (AUC) is then calculated and used as a single statistic measure. An AUC of 1 corresponds to a perfect discrimination between the classes (Zweig and Campbell, 1993), this is represented by the bold solid line in Figure 3.9 (b). Random discrimination is represented with an AUC of 0.5. If AUC < 0.5, then the predictions are negatively correlated with the ground truth. In this case, it is recommended to swap the class labels. This, however, should be applied to the entire dataset or not at all. The AUC is also known to be directly connected to the Mann-Whitney U-statistic which is a robust non-parametric alternative to the student's

t-test to assess the difference between two distributions. It has been extensively utilised in a number of classification tasks (Chai et al., 2017; Löfhede et al., 2008; O'Shea et al., 2017).

Accuracy is another widely used metric for the evaluation of classifier performance, which quantifies the percentage of correctly detected labels. This measure, however, relies on the particular threshold and reflects the performance of the classifier at a single coordinate of the ROC curve. Due to this limitation, in this thesis, only the area under the ROC curve statistic was employed.

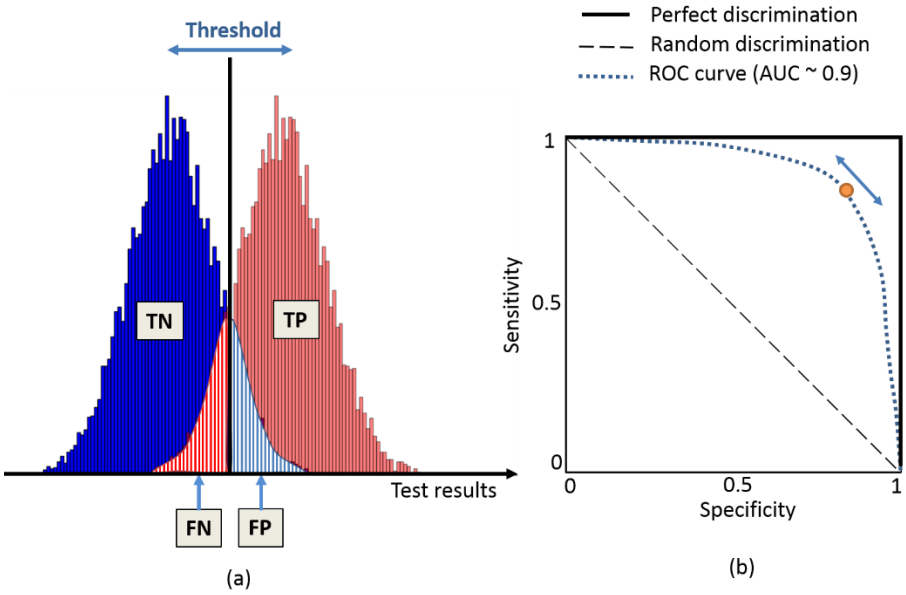


Figure 3.9: Description of the ROC curve and the area under the curve (AUC). Each point on the ROC curve represents a pair of sensitivity and specificity values according to the varying threshold.

3.4 Machine learning

ML is known as the field of study that allows computers to learn from the data without being explicitly programmed. In this sense, the term "learning" implies a progressive improvement of the performance on a specific task as well as the ability to generalise on previously unseen data. This is done by constructing a model based on the input training observations in order to generate a data-driven prediction of the output, rather than following predefined static program instructions.

3.4.1 Machine learning paradigms

The type of ML that learns a function which allows mapping an input data X to the output variable Y is called supervised learning. Unsupervised learning, on the other hand, is aimed at finding the patterns and structure in the data X without predefined variable or label Y. Other types of machine learning include reinforcement learning, where learning is performed via

interaction with an environment through trials and errors with the goal of long-term reward; active learning, which performs learning on a limited amount of training data and allows for the definition of training examples which are the most beneficial for a given problem.

Another categorisation of ML depends on the desired output that needs to be predicted. For the **classification** problem the labels are divided into two (binary classification) or more (multi-class classification) classes. This task is usually solved in a supervised manner. Outcome prediction is an example of the classification problem, where the output classes are 'healthy' and 'sick'. When the output of the system is a continuous value, rather than discrete a **regression** approach is used.

Clustering belongs to the unsupervised type of ML, in which the input data is divided into separate groups without the need for labels. This method is widely applied in different spheres: market segmentation, social network analysis, and others. In Google news, for instance, clustering is used for grouping news into cohesive stories. Clustering is also used in bioinformatics in order to build groups of genes with related patterns.

ML can be also used for the task of **dimensionality reduction**. This particular application is frequently used for data compression. More specifically, when dealing with a high number of features it is possible to simplify the input by mapping the features onto a lower dimensional space. Principal component analysis (PCA) is a well-known technique that allows for dimensionality reduction of the data while preserving the variance of the original data.

Generative and discriminative classifiers

The choice of the ML technique depends on the given problem. This thesis aims at predicting the health status of the preterm neonate at different stages of life based on the recorded physiological signals. For this problem, the supervised classification technique would therefore be the most appropriate. In order to build a model that predicts the response Y based on the explanatory variables (features) X, the dataset D is represented as $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$. To map every input $x \in X$ to a corresponding prediction $y \in Y$ a following function (called a classifier) is generated: $\hat{y} = f(x)$.

There exist two main approaches to construct the classifier: **generative** and **discriminative** approaches. Given the output Y and features X, a generative approach attempts to learn the joint probability distribution $P(X, Y)$, whereas the discriminative approach models the conditional probability of the label Y given the observations X, $P(Y|X = x)$. More specifically, the generative classifier models how the data was generated and then makes predictions using Bayes rule to produce the most likely output $\hat{y}$ as follows:

$$\hat{y} = \arg max_{y \in Y} P(y|x) = \arg max_{y \in Y} \frac{P(x|y)P(y)}{P(x)} \quad (3.25)$$

As it can be seen the generative classifier models $P(Y)$ and $P(X|Y)$, which are called the class prior and the class conditional distributions. Examples of the generative models are: naïve Bayes classifier, linear discriminant analysis, the Gaussian mixture model (GMM), Boltzman machine, autodecoders, adversarial NN and others.

A discriminative algorithm, on the other hand, makes no assumption on how the data is generated. In order to discriminate classes Y the discriminative approach directly learns the model $P(Y|X)$ depending only on the observed data. Examples of the discriminative classifiers are logistic regression, multilayer perceptron, SVM and decision trees. The discriminative classifier does not need to model the distribution of the observed data, and therefore it may not be able to express the possible complex relationship between observed variables and their labels. As a result, the discriminative models would not perform well on outliers. When the test data is generated by the different underlying distribution than the training data, it might be easier to tune the generative model according to the detected changes in its distributions. After fitting the generative classifier it is also possible to generate the data which is similar to the observed one. However, if the relationship between X and Y only approximates the true generative process, a discriminative model may be more preferred. In (Vapnik, 1999) the author argued that the classification problem should be solved directly by modelling $P(Y|X)$, without any intermediate steps. In practice, discriminative classifiers have been shown to outperform generative ones, especially for the cases when the number of training examples is high (Ng and Jordan, 2002). Therefore, if the main problem lies in the optimization of the classification accuracy and there is no necessity to make claims about the process of data generation, the discriminative classifiers may be well suited.

3.4.2 Supervised machine learning: decision trees

Decision tree algorithms are data-mining techniques which are used for both classification and regression problems. Using a tree-like model of the decision, these classifiers allow the user to visually and explicitly represent the decision-making process. All decision trees are constructed through recursion. A common tree structure can be defined as a root node followed by a set of internal nodes and final leaves. Each node is a logical divergent point where a particular explanatory variable (feature) splits the data according to a certain condition. All nodes are connected with branches showing the direction from a question to the answer. The leaf nodes are terminal nodes which have no child nodes and represent a value of a target variable. The schematic representation of a basic tree structure is shown in Figure 3.10.

A decision tree learns the decision boundary by recursively partitioning the input feature space. On each step, the algorithm chooses the feature which best splits the set of items according to a certain metric. The most common metrics are: Gini impurity, information gain, and variance reduction (Breiman et al., 1984). Gini impurity (I_G) measures the probability of a randomly chosen element being misclassified if the label was randomly selected according to the distribution of labels in a branch (Rokach and Maimon, 2005). It is calculated for every node as follows:

$$I_G = 1 - \sum_{i=1}^{J} p_i^2, \tag{3.26}$$

where p_i is the fraction of items labelled with class i in the given node. Ideally, two items which are randomly selected from the population should be of the same class and the probability of 1 corresponds to a pure population. Gini is the measure of impurity with zero value implying that the population contains one class only and no further split is needed. The information gain metric measures the expected reduction in entropy achieved after a split and is computed as follows:

$$Gain(S, A) = H(S) - H(S|A), \tag{3.27}$$

where S is a sample of training examples in the parent node; $H(S|A)$ is the entropy after the split on the feature A computed as a weighted sum of entropies in the two child nodes. The variance reduction metric is usually applied for regression problems. It uses the standard variance formula to choose the split with a lower variance.

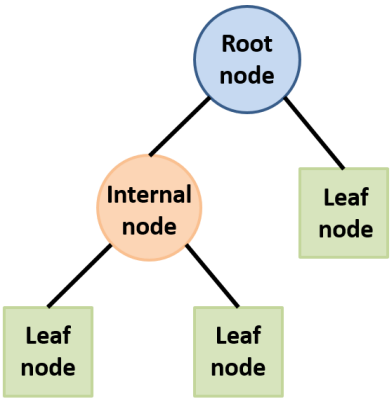


Figure 3.10: A basic structure of a decision tree.

Tree-based classifiers were previously applied for classification of the amplitude-integrated EEG of neonates in order to detect brain disorders (Chen et al., 2014). Simayijiang et al., (2013) used a tree-based classifier to study the EEG burst characteristics of preterm neonates. For the task of the EEG sleep stage classification, the decision trees outperformed SVM, NN, Naïve Bayes and K-nearest neighbours (Aboalayon et al., 2015). The study was conducted on

a single channel EEG from the publicly available dataset PhysioNet and resulted in classification sensitivity, specificity, and accuracy of 96.89%, 99.35%, and 97.30% respectively.

The decision tree model can be converted into a set of "if-then" rules, which makes it easy to interpret. In contrast to other "black box" modelling techniques, the main advantage of the tree-based classifiers lies in the possibility to find the reasoning behind the model. This property makes trees a good candidate for problems which require understanding of the decision-making process. While constructing decision trees only features which are useful for a given problem are included, which allows to use a tree-based classifier for the purpose of feature selection.

3.4.3 Supervised machine learning: support vector machine

The SVM is a discriminative classifier which uses a hyperplane in a high- or infinite-dimensional space to classify the data (Vapnik, c1982.). In order to discriminate the data which is not linearly separable SVM maps the data onto the higher dimensional feature space, where the data will become more linearly separable as compared to the input feature space. The mathematical function used to transform the data is called a similarity function or a kernel. Once the data is mapped it can be separated using a hyperplane. The SVM is also known as a maximum margin classifier, which is due to the way the SVM builds its decision boundary. An illustration of the SVM decision boundaries for two perfectly separable classes is represented in Figure 3.11. For this linearly separable data two supporting hyperplanes are selected so that the distance between them, called the margin, is as large as possible. The best separating hyperplane is the one that lies halfway between the supporting hyperplanes.

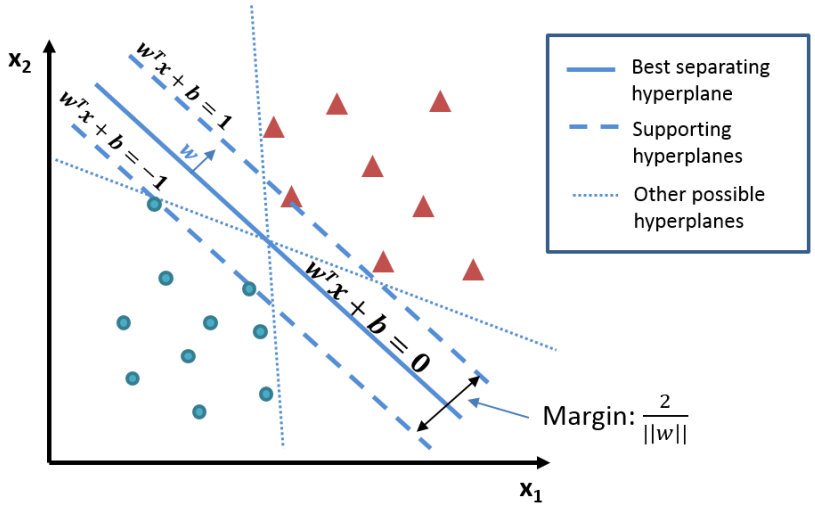


Figure 3.11: SVM decision boundary for linearly separable class.

For a dataset of N points in the form of feature vectors $\{x_1, x_2, \dots, x_N\}$ with corresponding labels $\{y_1, y_2, \dots, x_N\}$, where $y_i \in \{-1,1\}$ a hyperplane is defined as: $w^T x + b = 0$ with w corresponding to the normal vector to the hyperplane; $\frac{|b|}{||w||}$ is a perpendicular distance from the hyperplane to the origin; $||w||$ is the Euclidian norm of w. As it can be seen (Figure 3.11) there also exist other possible hyperplanes which can perfectly separate the two classes. However, none of them provide the maximum possible distance between two classes of data. The distance that has to be maximised is $\frac{2}{||w||}$, which implies minimising $||w||$. The final SVM classification decision is obtained according to the following function:

$$f_{svm}(x) = sign\left[\sum_{i=1}^{N} \alpha_i y_i k(x_i, x_j) + b\right], \tag{3.28}$$

where k is a kernel (similarity) function for features x_i and x_j; $\alpha_i \geq 0, i = 1, \dots, N$ are the Lagrange multipliers; the *sign* function determines whether predicted value comes from the positive or negative class. The most popular kernel is the Gaussian radial basis function (Aslan et al., 2008; Li et al., 2014) defined as:

$$k(x_i, x_j) = \exp\left(\frac{-||x_i - x_j||^2}{2\sigma^2}\right) \tag{3.29}$$

The SVM is widely used both in industry and academia for classification and regression problems. SVM showed state-of-the-art results in the field of EEG-based neonatal seizure detection with a detection rate of 96% with two false detections per hour (Temko et al., 2011). It was also successfully applied for the grading of the hypoxic-ischemic encephalopathy in neonates EEG (Ahmed et al., 2016).

3.4.4 Supervised machine learning: Gaussian mixture model

The GMM is a generative probabilistic model. It assumes that all data points which represent the feature space can be modelled by a mixture of Gaussian distributions with unknown parameters (Reynolds, 2009). The GMM can be represented as a weighted sum of *M* Gaussian components. The likelihood of the data given class *c* is defined as follows:

$$p(x|c) = p(x|\theta_c) = \sum_{j=1}^{M_c} w_j^c g(x|\mu_j^c, \Sigma_j^c), \tag{3.30}$$

where $\theta_c = \{w_1^c \dots w_{M_c}^c, \mu_1^c \dots \mu_{M_c}^c, \Sigma_1^c \dots \Sigma_{M_c}^c\}$, x is a feature vector of n dimensions; w_j are the mixture weights with $j = 1, \dots, M$; $g(x|\mu_j, \Sigma_j)$ are the Gaussian component densities, each described as:

$$g(x|\mu_j, \Sigma_j) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)\right), \quad (3.31)$$

where μ_j and Σ_j are the mean and covariance matrices of the j^{th} Gaussian component. An example of the GMM with two components is represented in Figure 3.12.

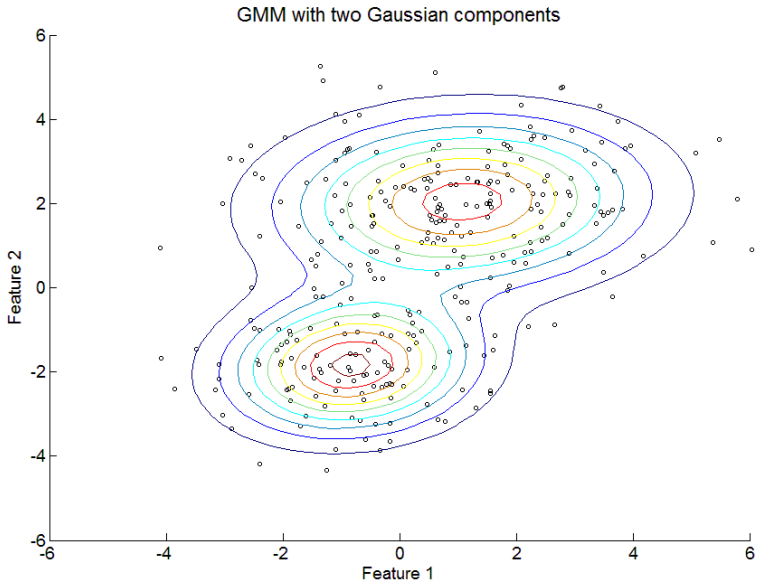


Figure 3.12: An example of the Gaussian mixture model with two Gaussian components generated for one class of data.

During training, the parameters θ_c are optimized in order to maximize the likelihood $p(X|\theta_c)$ of the training data for each class individually. The most common algorithm is the expectation maximization (EM), which starts with some initial parameters θ_c, and iterates to find a new set $\widehat{\theta_c}$ with a constraint $p(X|\widehat{\theta_c}) \geq p(X|\theta_c)$. This iterative process repeats until a certain stop condition defined according to the number of iteration or a convergence threshold. The probability of data X belonging to component C_c out of all the n_c possible classes, is defined using Bayes' theorem:

$$P(C_c|X) = \frac{p(X|\theta_c)}{\sum_{k=1}^{n_c}(X|\theta^k)}, \quad (3.32)$$

GMMs were extensively used for the problems of seizure classification (Thomas et al., 2009) and artefact detection (Kauppila et al., 2018) using neonatal EEG signals. A study (Costa et al., 2012) analysed HRV using GMM and reported that HRV can be predicted by the mixture of three Gaussian components. The authors suggested that different components could identify the activity of different branches of the adult autonomic nervous system.

3.4.5 Supervised machine learning: neural networks

The neural network (NN) model is a system which was inspired by the biological neural network and mimics the brain function. Nowadays, NN is the state-of-the-art technique for many supervised ML problems. Just like the brain, a NN is composed of neurons, whose outputs generate nonlinear functions from its input signal. An input signal parameterised by weights travels to the output of the network through a number of layers, where each layer performs a different type of nonlinear transformation of the input signal. A schematic representation of typical multilayer NN is shown in Figure 3.13.

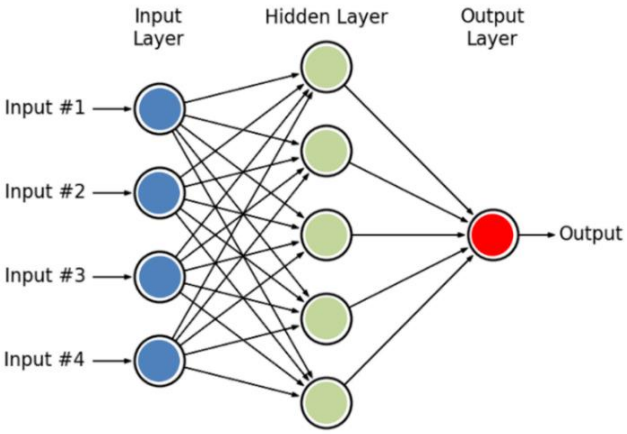


Figure 3.13: A diagram of the feedforward neural network with one hidden layer.

The main idea of the NN is that the weights *w* are learnable and are able to control the influence of one neuron on another. As it can be seen from Figure 3.14 the weight vector input of the neuron is summed. If the sum is above some definite threshold, the neuron can fire, which is modelled with a nonlinear activation function σ. Each neuron performs a dot product of the input and its corresponding weights, adds a bias term (which shifts activation function to allow for better learning) and then applies the nonlinearity with different activation functions (Figure 3.15). Therefore, rather than using original features, the NN uses new learned features, which are functions of the input. On each level of a multi-layered NN, the input data is transformed into a more abstract and composite representation. Learning through a cascade of multiple layers of nonlinear processing elements is referred to as deep learning (DL).

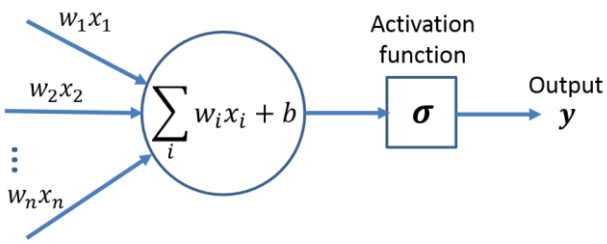


Figure 3.14: An example of a neuron showing the input and its corresponding weights and the activation function applied to the weighted sum over the inputs from the previous layer.

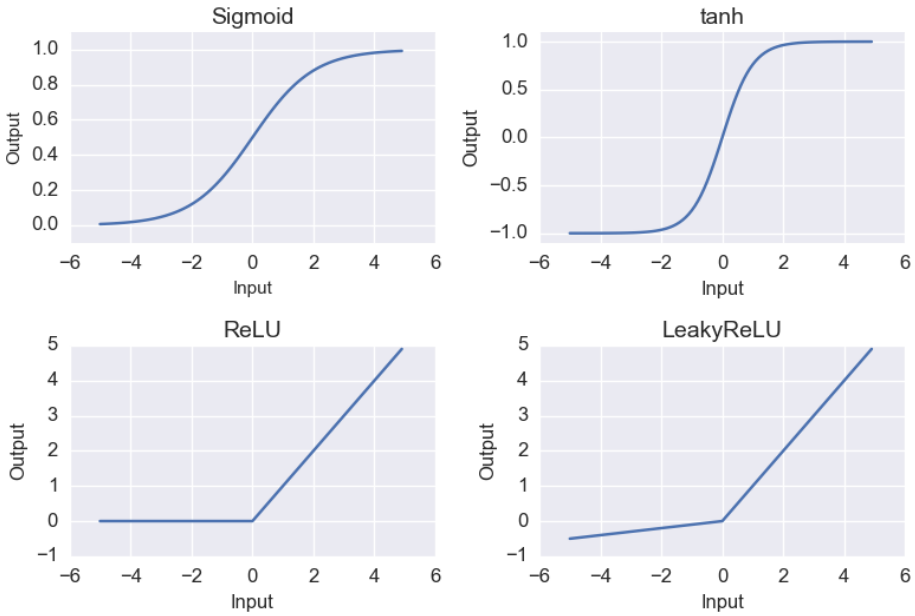


Figure 3.15: Examples of activation functions.

The training process of the NN starts with the random initialisation of the network parameters w. This allows for the breaking of the symmetry and enables hidden nodes to learn different features. In order to check the performance of the initial hypothesis of the network using the initial parameters w and input x, forward propagation is performed.

During the process of training (backpropagation) the accuracy of prediction of the model is improved, or some specific loss function which evaluates how well the algorithm models the given data is minimized. The choice of loss function depends on the given task. For regression problems the most common losses are mean squared error (L2 loss) and mean absolute error (L1 loss). Given the classification task, the most common loss functions are the hinge loss, which is usually used for SVM with the desired output represented as $\{+1, -1\}$ and log loss (categorical cross-entropy loss) where the output is represented as $\{0, 1\}$. The log loss increases as the predicted probability diverges from the label; it is defined as:

$$L(y, \hat{y}) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \tag{3.33}$$

Here L is a loss function that measures the difference between the prediction $\hat{y}_i$ and the target y_i; p_i is a logistic function defined as: $p_i = \frac{1}{1+e^{-\hat{y}_i}}$. This function measures the performance of the model where the output is a probability value defined between 0 and 1. Figure 3.16 represents the visualization of the logarithmic loss function when the true label is equal to 1. It can be seen that as the predicted probability approaches the true label, the log loss decreases.

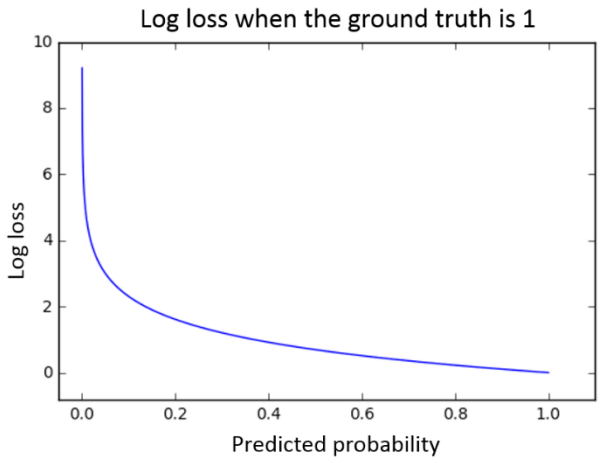


Figure 3.16: Visualisation of the possible log loss values given a ground truth of 1.

The optimization of a loss function can be done using different techniques, with the gradient descent algorithm known as the most common one. The main task of gradient descent is to iteratively update the parameters w with a certain learning rate α until convergence. This is achieved by computing partial derivatives $\partial L/\partial w$ and $\partial L/\partial b$ of the loss function L with respect to weights w and bias term b. A single iteration of the gradient descent updates the parameters in layer *l* as follows:

$$w_{ij}^{(l)} = w_{ij}^{(l)} - \alpha \frac{\partial L}{\partial w_{ij}^{(l)}} \tag{3.34}$$

$$b_{i}^{(l)} = b_{i}^{(l)} - \alpha \frac{\partial L}{\partial b_{i}^{(l)}} \tag{3.35}$$

The backpropagation algorithm performs the backward propagation of errors, which allows for the computation of the partial derivatives by applying the chain rule. During this procedure, the slope of the derivative of activation function is calculated. Sometimes, when training deep networks, the gradients can become very small and lead to the vanishing gradient problem. As it can be seen from Figure 3.15, on the regions where the slope of the functions (sigmoid and tanh) is close to zero, the learning process may become very slow. The rectified linear unit (ReLU) activation function has partially addressed this problem of vanishing gradients. For the ReLU the gradient is equal to one for all positive input and therefore cannot shrink to zero as the neuron saturates. This function is not computationally expensive and is easy to use during backpropagation. ReLU activation simply sets a threshold at zero according to:

$$\sigma(x) = \begin{cases} 0 \; if \; x < 0 \\ x \; if \; x \geq 0 \end{cases} \tag{3.36}$$

The leaky ReLU is an improved version of ReLU. It incorporates δ, a small constant usually equal to 0.01, which helps to increase the range of ReLU function. Leaky ReLU is computed as follows:

$$\sigma(x) = \begin{cases} \delta x \; if \; x < 0 \\ x \; if \; x \geq 0 \end{cases} \tag{3.37}$$

Types of gradient descent

There are different types of gradient descent, which mainly differ in the amount of data they use. The most common one is batch gradient descent, which calculates the error for each training example within a dataset and updates the model over the whole batch. A single iteration of this process is usually called a training epoch. This type of gradient descent generates a stable error which is based on all training examples. However, this technique might be challenging to use, particularly in cases where the number of training examples is too high and it is not possible to allocate enough memory for the entire dataset. Stochastic gradient descent (SGD), on the other hand, performs a model update for each training example, which can result in noisy gradients and therefore complicate the process of error decrease. Mini-batch gradient descent represents a trade-off between the previous two methods. This technique splits the dataset into smaller batches and performs the model update for each mini-batch. This encourages smoother gradients and also allows for the selection of batch size suitable for the amount of memory available.

Variants of neural networks

There exist different types of NN, such as convolutional NN (CNN), long short-term memory networks (LSTM), recurrent neural network (RNN), residual NN (ResNet) and others. Very deep NN may be difficult to train due to the possible problem of either the vanishing or the exploding of the gradients (Bengio et al., 1994; Hanin, 2018). The winner of the ILSVRC 2015 classification task proposed deep residual learning using a 'residual network' known as ResNet, which addressed the problem of training deep networks by reusing activation from the previous layers (He et al., 2015a). The main concept of ResNet is represented in Figure 3.17.

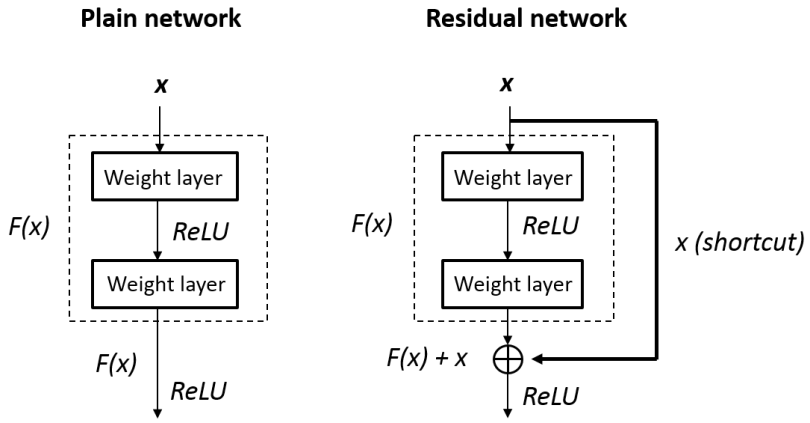


Figure 3.17: The main concept of conventional and residual networks.

ResNet incorporates so-called skip connections (shortcut), which facilitate the learning of much deeper NN by learning the residuals of input and output of some layers. During backpropagation, the gradient is propagated without modifications directly to the earlier layers using skip connections.

3.4.6 Convolutional neural network

The CNN is yet another feedforward NN originally developed for images analysis as a special case of the traditional multilayer perceptron. A large number of pixels (features) typically present in an image creates a significant challenge for a conventional densely connected NN, as it requires a large number of trainable parameters. This may eventually create a problem of overfitting, unless a sufficiently large training dataset is available, as well as make the training procedure not feasible due to high computational and memory requirements. The CNN deals with the spatial variance of features and a large number of pixels by making some explicit assumptions that the inputs of the network are images. This is done by introducing filter matrices which are moved across a 2D image. The multiplication of the filter matrix with the input layer represents the correlation between them. The result of mapping (correlation) is then passed to the next layer. This operation is known as a convolutional operation.

Feature maps

Images typically consist of edges and lines that hierarchically form more complex shapes and objects. Element-wise product of the filters and the data during the convolution operation performs the edge detection in the first layers and extraction of more complex features in the following layers. Examples of the horizontal and vertical edge detectors are represented in Figure 3.18. In can be seen, that a vertical edge detector, for instance, looks for the bright pixels on the left and dark ones on the right, while ignoring the middle. Learning different filter parameters will eventually allow learning different edges as well as many other filters.

1	0	-1
1	0	-1
1	0	-1

1	1	1
0	0	0
-1	-1	-1

Figure 3.18: Examples of the filters which perform the vertical and horizontal edge detections.

The mapping quantifies the correlation of the feature with each part of the input and is referred to as a feature map. It is obtained by sliding the filter across the width and height of the input and performing the dot product between the filter and input at each spatial location. Depending on the number of channels at the input (e.g. 3 channels representing RGB colours in images), the filter can have an additional dimension for depth, resulting in 3D filter size. For the input image I and a filter $W \times H \times C$ the output of convolution operation can be defined as follows:

$$x_{i,j}^{l} = b^{l} + \sigma\left(\sum_{c=0}^{C}\sum_{w=0}^{W}\sum_{h=0}^{H} w_{w,h,c}^{l}\, x_{(i+w),(j+h),c}^{l-1}\right) \tag{3.38}$$

Where $x_{i,j}^{l}$ is a convolved input vector at the layer l; i and j are the indices for the dimension of the input I; $\sigma(.)$ is an activation function applied to the convolved input at the layer l; w is a filter of weighs and b is a bias term.

The depth of the convolutional filter matches the depth of the input, where the filter at each depth has different weights. The convolution operation performed on a 3D filter will still result in a scalar obtained by multiplication and summation of feature maps as shown in Figure 3.19. Therefore, when using five different filters, for instance, the convolution for each of them would be performed independently and a final result would generate five feature maps stacked along the depth dimension. The sharing of parameters (weights) across the 2D input allows the network to search for the same pattern in all locations. This assumes that if some feature is useful in one location, it may also be useful and reused in other locations as well. This parameter sharing scheme of the CNN provides translation invariance to the algorithm and also reduces the number of parameters.

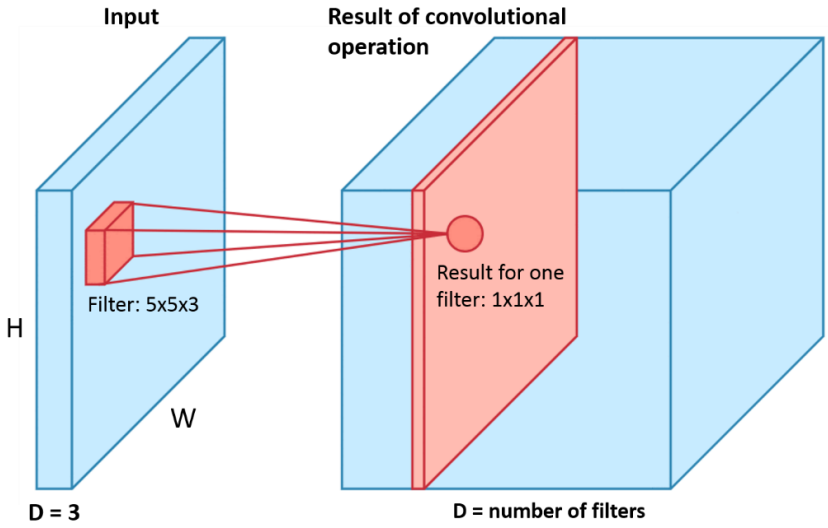


Figure 3.19: Schematic representation of the convolutional operation for a 3D input. The size of the filter is 5x5x3, where each depth of filter is represented by a different matrix of weights. The convolution performed on the 3D input, results in a scalar obtained by multiplication followed by summation.

Spatial arrangements

There are few additional hyperparameters which regulate the output dimension of the convolutional layer. During the convolutional operation, the pixels on the edges of the image are only mapped once by each filter, whereas pixels away from the edge are mapped multiple times. As a result, not all information on the edges is fully used. Zero-padding of the input

performed prior to the convolutional operation allows to avoid this problem, as well as to prevent shrinkage of the input. Stride is another parameter which regulates the number of pixels to move the filter at a time. Higher values of stride allow for faster downsampling of the input image, if necessary. As a result, depending on the stride s and padding p parameters, the output dimension n_{out} of the convolution operation can be defined as follows:

$$n_{out} = \frac{n_{in} + 2p - f}{s} + 1 \tag{3.39}$$

where f and n_{in} are the dimensions of filter and input correspondingly.

Pooling

The pooling layer does not perform any learning and is responsible for reducing the height and the width of the input. This operation partitions the input into regions and outputs the maximum (max pooling) or average (average pooling) for each of them (Figure 3.20). This layer operates separately on each depth slice. Pooling performs nonlinear downsampling which reduces the number of parameters. It allows to avoid overfitting, as well as to handle inputs of varying sizes. The pooling operation also contributes to translation invariance, by indicating the presence of feature rather its exact position.

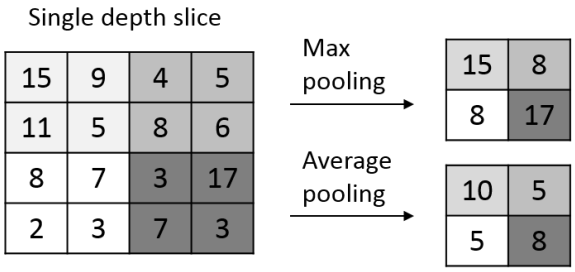


Figure 3.20: An example of the max and average pooling operation with a 2x2 filter and a stride of 2.

Receptive field

The receptive field is a region in the input space by which each convolution is influenced. More specifically, the first convolutional operation results in a receptive field which is equal to the filter size. In the deeper layers, the receptive field gets wider. Most commonly, the term receptive field is referred to the widest receptive field in the final convolutional layer in relation to the network input. The size of the receptive field can be increased by adding more convolutional layers (depth), downsampling (pooling, stride) and using filter dilation (Yu and Koltun, 2015).

Batch normalization

During the training procedure the parameters in each layer change, and as a result, the distributions of the inputs of each layer change as well. This phenomenon is referred to as internal covariance shift (Ioffe and Szegedy, 2015). It was previously shown that a network converges faster if its inputs are normalised (Wiesler and Ney, 2011). The batch normalisation (BN) technique which is used in the CNN ensures that inputs to the subsequent layers are normalized by scaling the activations to have zero mean and unit variance. The level of normalization is adjusted during the training procedure by learning two additional parameters which scale and shift the normalized values of each activation. This technique is applied before the nonlinear activation function and allows the use of higher learning rates which can then speed up the training procedure. The algorithm of batch normalization (Ioffe and Szegedy, 2015) for a particular activation x applied over a mini-batch B to generate y is represented below.

Algorithm 3.1: Batch normalization transform

Input:

Values of x in a mini-batch: $B = \{x_{1,} x_{2,} \dots x_m\}$;

Learnable parameters: γ, β

Output: $\{y_i = BN_{\gamma,\beta}(x_i)\}$:

$\mu_B \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$ // mean of mini-batch

$\sigma_B^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_B)^2$ // variance of mini-batch

$\widehat{x_i} \leftarrow \frac{x_i - \mu_B}{\sqrt{\sigma_B^2}}$ // normalization

$y_i \leftarrow \gamma\widehat{x_i} + \beta \equiv BN_{\gamma,\beta}(x_i)$ // shift and scaling

Regularization techniques

One of the main aspects of training a NN is maintaining the bias-variance trade-off. While capturing the specifics of the training data it is important to be able to generalise to unseen data as well. Overfitting (high variance) is a very common problem which occurs when the model captures noise in the training data. There exist several regularisation techniques which can be deployed in the NN in order to avoid overfitting.

The training procedure is based on the learning of network parameters (weights) which allow to minimise a defined cost function. In order to reduce the variance in the network performance, the regularisation parameter can be explicitly added into the cost function. The most common type of regularization is L2 regularization, which aims at minimising the weights of the network. A cost function J can be regularised in the following way:

$$J(w^1, b^1, \dots, w^L, b^L) = \frac{1}{2}\sum_{i=1}^{m}\left(\hat{y}^{(i)}, y^{(i)}\right) + \frac{\lambda}{2m}\sum_{l=1}^{L}||w^l||_2^2 \tag{3.40}$$

where λ is a regularization parameter; $||w^l||_2^2$is a squared 2 norm of the weight matrix. It can be seen that higher values of λ will result in smaller weights. This regularisation parameter can be tuned during the CV procedure.

Overfitting can also be tackled by reducing the complexity of the network. A reduced number of parameters can be achieved by using fewer filters and /or hidden layers (i.e. network depth). The alternative would be to use the same classifier architecture but with more training data. If possible, data can be generated from the scratch or augmented from the existing training set. An early stopping technique is yet another way to prevent overfitting, which stops training before the overfitting has occurred. Dropout is an additional regularisation technique which performs random dropping of activation units and their corresponding input and output connections from the network during the training procedure (Srivastava et al., 2014). Dropout can be also viewed as a sampling of a variety of smaller networks from the original larger one.

Unlike the standard feedforward NN, LSTM and RNN can solve problems of temporal dependencies by maintaining information in "memory". LSTM is a type of RNN which allows for leaning the long-term temporal dependencies (Hochreiter and Schmidhuber, 1997). These networks have been successfully applied to sequential data, such as audio and text, for problems of speech recognition, handwriting recognition, and others. Despite the fact that the RNN is usually the starting point for sequence modelling, recent research has demonstrated that certain convolutional architectures are able to reach state-of-the-art performance in machine translation and audio-synthesis (Dauphin et al., 2016; Oord et al., 2016). Bai et al., (2018) conducted an empirical evaluation of generic recurrent and convolutional architectures for the sequence modelling. New architectural elements of the CNN, such as residual connections; causal convolutions (when the output at time *t* only depends on the data before *t*); and dilated convolutions which increase the receptive fields, allowed these networks to outperform recurrent architectures. Authors have concluded that convolutional networks may have longer memory and therefore should be considered as a starting point for the sequence modelling task.

The NN is a very powerful technique which has outperformed human benchmarks in image processing tasks (He et al., 2015b). The CNN applied to raw EEG showed the state-of-the-art result in the field of neonatal seizure detection. It has outperformed the previous state-of-the-art SVM algorithm that was based on 55 hand-crafted features (O'Shea et al., 2017). Another study applied CNN to EEG signals in order to decode and visualise brain pathology (Schirrmeister et al., 2017). NNs were also applied to other physiological signals, including HRV and BP signals for the tasks of an accurate heartbeat detection (Bollepalli et al., 2018), assessing driver fatigue based on heart activity (Patel et al., 2011) and others.

3.5 Bias-variance trade-off in machine learning

In supervised ML the algorithm learns a model from the given training data. As mentioned earlier, the main goal is to estimate some function $f(X)$ which maps the input data X onto the output variable Y. For any type of machine learning technique, there are three main parts of the prediction error: variance error, bias error, and unavoidable bias (optimal error rate /Bayes error rate). In general terms, the bias is the difference between the expected and true value of the estimator f:

$$Bias = E\left[\hat{f}(X)\right] - f(X) \tag{3.41}$$

When this difference is equal to zero, then $\hat{f}$ is referred to an unbiased estimator of f. Models with a high bias oversimplify the problem when mapping X onto Y and lead to a high error on the training and test data. The error rate of the algorithm on the training set is known as bias. The variance error computes the variance of the estimator $\hat{f}$ and its expected value:

$$Variance = E\left[(\hat{f}(X) - E[\hat{f}(X)])^2\right] \tag{3.42}$$

The variance estimates the variability of the predictions if the learning process is repeated multiple times with small changes in the training set. Algorithms with high variance are strongly influenced by the specifics of the training data and as a result, are not able to generalise on previously unseen data. These models perform very well on the training set but have a high error on the validation /test sets. Usually, the data contain noise and the problem of mapping X onto Y can be represented as $Y = f(X) + e$, where e is the lowest possible prediction error or optimal error which cannot be avoided. The total error is represented as a sum of all errors:

$$Total\ error = \left(E\left[\hat{f}(X)\right] - f(X)\right)^2 + E\left[(\hat{f}(X) - E[\hat{f}(X)])^2\right] + e^2 \tag{3.43}$$

The visual representation of bias and variance problems and the trade-off between the two is shown in Figure 3.21.

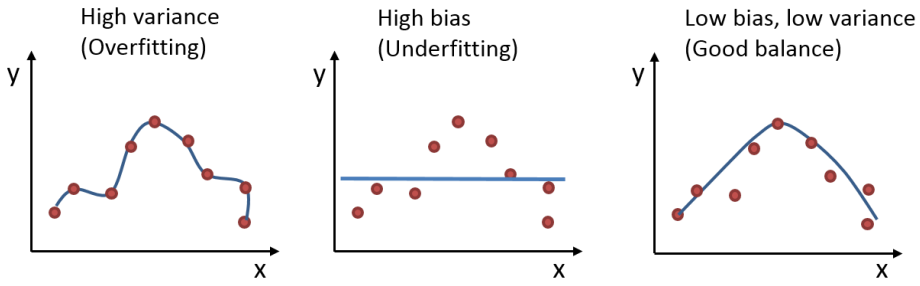


Figure 3.21: Visual representation of the bias-variance trade-off in machine learning.

The bias-variance trade-off is the main problem in supervised ML. Ideally, the classifier should be able to capture the specifics of the training data and at the same time generalise to the unseen data of the test set. Different ML techniques are prone to either of these problems. More specifically, decision trees, SVM and k-Nearest Neighbours are typical examples of the low bias and high variance problems, whereas linear and logistic regressions are known as low variance, high bias algorithms. The main goal of any algorithm is to achieve the bias-variance trade-off, which is visually represented in Figure 3.22. This is usually done with a parameterization which regulates the complexity of the model.

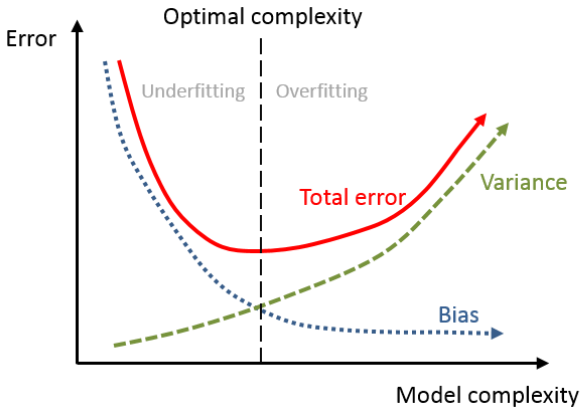


Figure 3.22: The bias-variance trade-off represented with learning curves. The total error represents the generalisation error of the model.

In order to properly tune the model, first, it is necessary to investigate the performance over the training set. If the bias problem (underfitting) is identified, it is necessary to increase the complexity of the model (larger network, more iterations for NN or a deeper tree for decision trees) since the current mapping function $f(X)$ is too simple to predict Y. Once the classifier is trained on the training set, the performance of the validation set is assessed. If we find ourselves in the region or high variance (overfitting) it is possible to apply some regularisation techniques (L1, L2 regularizations, dropout in NN etc.) or use more training data.

When training the classifier it is also important to make sure that its performance is properly evaluated. The optimal model is chosen as the one which performs the best on the validation

set. After the model is selected, the performance is estimated on the held out test set. This procedure is schematically represented in Figure 3.23.

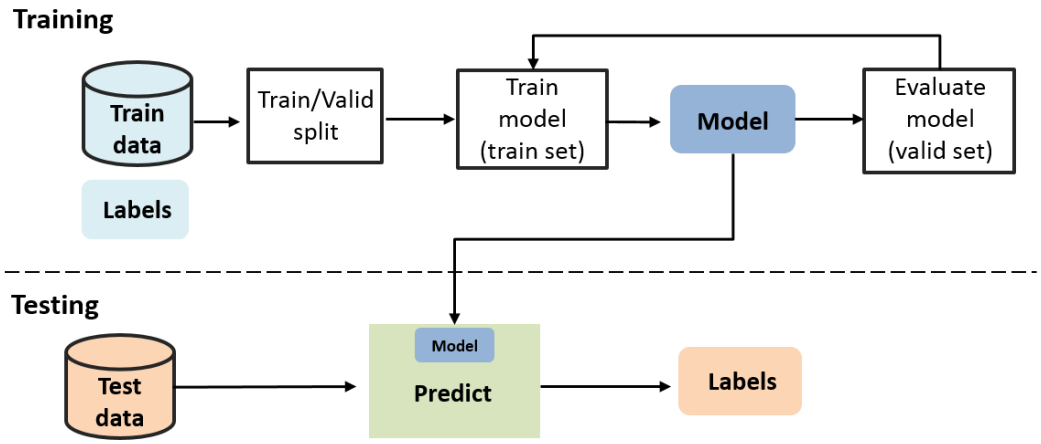


Figure 3.23: Schematic representation of the supervised machine learning workflow.

3.6 Conclusion

This chapter presented a description of signal processing and feature extraction techniques used in the thesis. The measures of linear and nonlinear interaction employed for the quantification of coupling between physiological signals were discussed. The chapter also provided a general overview of the different ML techniques along with their applications in the medical field. Given the objectives of this thesis, the provided information will allow for the selection of suitable tools for each individual problem investigated.

Chapter 4: Modelling interaction between blood pressure and brain activity

This chapter will provide a detailed explanation of the interaction modelling between brain activity and BP, and its association with the clinical risk index for babies, which characterises the health status of the preterm neonate soon after birth. The level of coupling between the two physiological systems was estimated using linear and nonlinear methods such as correlation, coherence and mutual information. Transfer entropy is also computed, to provide insight into the directionality of the interaction between EEG and BP. This detailed analysis allows for a comprehensive picture of the available data. The reliability of the obtained results was checked by testing an appropriate null hypothesis for every computed measure of interaction using surrogates.

4.1 Introduction

Hypotension or low blood pressure (BP) is a common problem in preterm neonates and criteria which defines hypotension is still not clear (Dempsey and Barrington, 2007). Deciding when and whether to treat hypotension relies on our understanding of the relation between BP, oxygenation and brain activity. However, little is known about this relationship in preterm infants as these signals are rarely recorded simultaneously and the extraction and investigation of the complex measures of signal interaction and signal dynamics have not yet been explored. In this chapter, we aim to investigate the interaction (coupling) between BP and continuous multichannel unedited EEG recordings in preterm infants less than 32 weeks GA. Detection of relationships and the quantification of interactions between physiological systems can be carried out in different ways, with classical linear methods, coherence, and correlation being the most common (Shaw, 1981). These measures capture linear relationships only and therefore fail to detect nonlinear coupling between the signals.

This study hypothesises that a nonlinear measure of interaction between BP and EEG may be more sensitive to adverse health conditions than linear methods. In this work, the linear interaction between EEG and BP is quantified by classical coherence and correlation measures, while the nonlinear coupling is computed based on MI. Hypothesising that

difference in coupling is indicative of the level of preterm wellbeing, we test the association between this dynamic coupling and the illness risk score.

While other studies mainly analysed preselected short EEG epoch, the main strength of the current work lies in the analysis performed on long duration unedited multichannel EEG recordings (total of 957 hours).

4.2 Dataset

For this study, the wellbeing of the preterm neonate was evaluated using the clinical risk index for babies (CRIB II) (Parry et al., 2003). Compared to other risk scores, CRIB II (further referred to as CRIB) is found to have an improved discriminative power when assessing mortality risk for VLBW infants (<1500g) (Gagliardi et al., 2004). From Figure 4.1 it can be seen that CRIB score is assigned soon after birth (within the first 12 hours of life) and provides an early diagnosis of the mortality risk of the preterm (Parry et al., 2003).

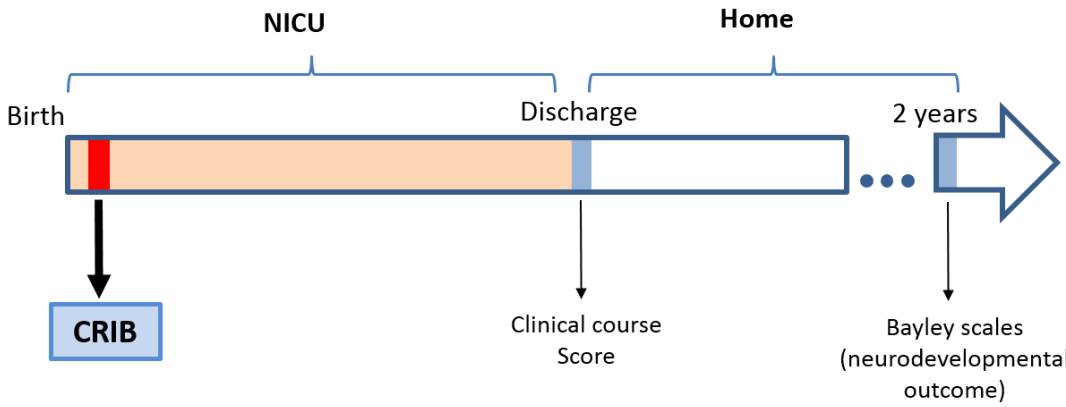


Figure 4.1: The timing of CRIB score along with other illness scores considered in this thesis are assigned to an infant during the course in the neonatal intensive care unit (NICU) through to the neurodevelopmental follow-up at 2 years of age.

The analysis is performed on a database of EEG data from 25 preterm infants < 32 weeks GA (range: 23 – 31 weeks) with available CRIB score. The data were recorded at the NICU of Cork University Maternity Hospital, Ireland. Clinical characteristics are provided in Table 4.1. The dataset included continuous multichannel EEG, simultaneous registration of BP and CRIB scores. The duration of recordings used in this study totals 957 hours (median = 37 hours, IQR = 24 to 48 hours). Figure 4.2 represents the duration and temporal location of each recording with the time of birth as a reference point. It can be seen that all recordings were initiated soon after birth.

The EEG data were sampled at 256 Hz (21 subjects) and 1024 Hz (4 subjects). BP was sampled at 1 Hz. An example of a simultaneously recorded EEG and BP data segment is shown in Figure 4.3.

Table 4.1: Clinical information represented as median (IQR).

Subject #	GA (weeks)	BW (g)	CRIB	Apgar score 5 min	Gender	Umbilical cord pH	Recording duration (hours)
1	24	670	17	6	M	6.85	55
2	24	740	13	9	F	7.18	48
3	23	540	14	7	F	7.22	49
4	28	1040	7	9	F	7.3	24
5	25	730	12	7	F	7.29	43
6	30	1450	6	5	M	7.02	68
7	26	950	12	6	F	6.96	39
8	31	960	7	9	F	7.23	37
9	25	620	12	6	F	7.34	39
10	26	860	10	8	M	7.12	24
11	26	980	10	8	M	7.24	39
12	30	730	10	10	M	7.32	24
13	24	1240	5	9	F	7.15	24
14	28	1330	4	4	F	7.08	24
15	30	1000	7	10	F	7.24	24
16	28	650	9	7	F	7.22	43
17	28	980	8	8	F	7.16	24
18	28	1060	6	1	M	6.9	83
19	29	1230	7	9	M	7.27	37
20	24	620	13	9	F	7.27	24
21	30	800	10	8	M	7.28	28
22	28	980	8	10	F	7.26	37
23	30	1530	3	8	F	7.1	24
24	23	580	14	6	F	7.24	48
25	28	680	8	8	M	7.32	48
Median (IQR)	28 (25 to 29)	950 (680 to 1040)	9 (7 to 12)	8 (6 to 9)	64% (F)	7.23 (7.12 to 7.27)	37 (24 to 48)

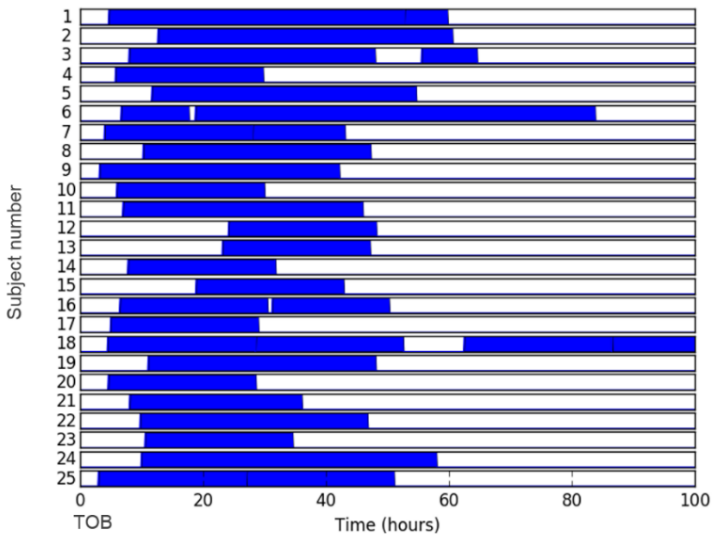


Figure 4.2: Schematic representation of duration and temporal location of recordings. Each recording is represented with respect to the time of birth (TOB) for each neonate.

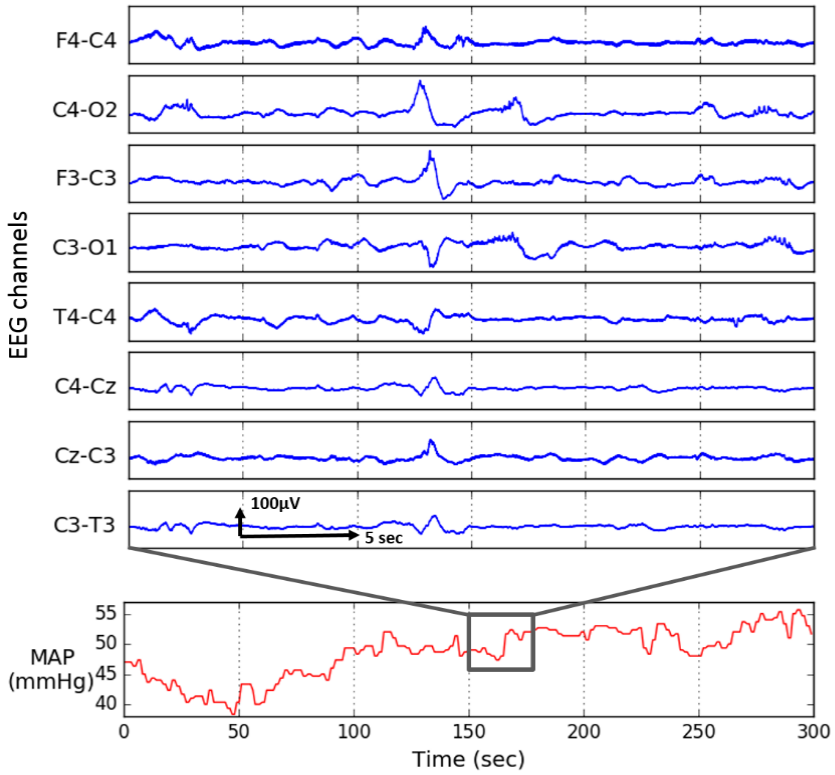


Figure 4.3: Five minutes of mean arterial pressure (MAP) and 30 seconds of 8-channel raw EEG recording (30 weeks GA).

Artefacts in both EEG and BP were detected and removed using the pre-processing techniques explained in Chapter 3. This study had full ethical approval from the Clinical Research Ethics Committee of the Cork Teaching Hospitals. Parental written informed consent was obtained for all newborns recruited for EEG monitoring studies. All data were anonymised.

4.3 Exploring the contextual information of signals and features

EEG

The brain of the preterm neonate is different to the adult brain and is characterised by complex spatiotemporal information (Pavlidis et al., 2017). Proper interpretation of neonatal EEG requires knowledge of the brain developmental changes from early preterm to post-term age. In order to account for all the maturational changes happening within a brain of the preterm, four main sub-bands (0.3-3 Hz, 3-8 Hz, 8-15 Hz, and 15-30 Hz) of electrical cortical activity were considered in this study. An example of EEG waveforms of the preterm neonate in the different frequency sub-bands is represented in Figure 4.4. The 0.3-3 Hz (delta activity) component is a slow high amplitude activity and it is known to be a common background component of the preterm EEG.

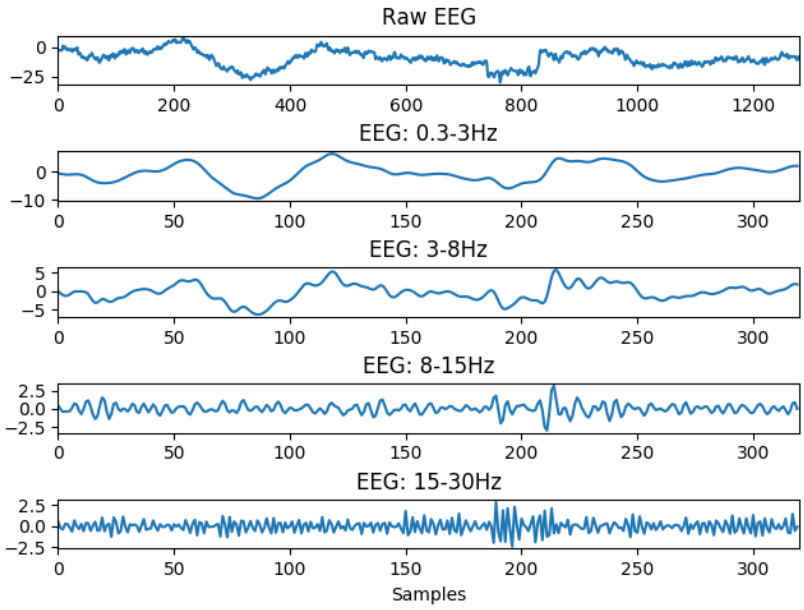


Figure 4.4: An example of a 5-second trace of F4-C4 channel of raw EEG (fs=256 Hz) and its corresponding waveforms in different sub-bands (fs=64 Hz) of preterm with GA of 27 weeks.

In order to measure the coupling between physiological signals, it is necessary to derive the informative values (features) that characterise the measured data. Prior to EEG feature extraction, the EEG signal is filtered to the range of 0.3-30 Hz and down-sampled to 64 Hz. The EEG is then segmented into 1-minute epochs with the 1-second shift. The EEG was quantified by the power in each of four frequency bands (0.3-3 Hz, 3-8 Hz, 8-15 Hz, and 15-30 Hz) as explained in Chapter 3. An example of the EEG spectrum obtained using the Fourier transform is represented in Figure 4.5. As expected, it can be seen that most power of the preterm EEG in concentrated in the low frequencies of delta sub-band (0.3-3 Hz). It was previously reported (Bell et al., 1991) that delta activity in the preterm EEG (<32 weeks GA) accounts for about 80% of the relative power for frequencies below 1 Hz. Each EEG feature is summarised as a median across eight bipolar channels. An example of the resultant traces of four EEG sub-band powers is represented in Figure 4.6.

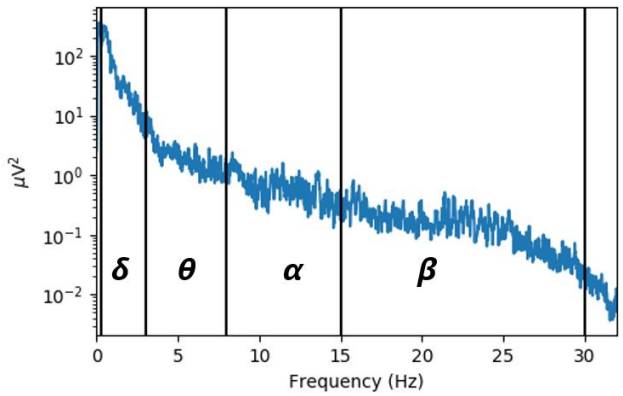


Figure 4.5: EEG spectrum of a preterm neonate (27 weeks GA) computed over a 10-minute trace of the F4-C4 channel.

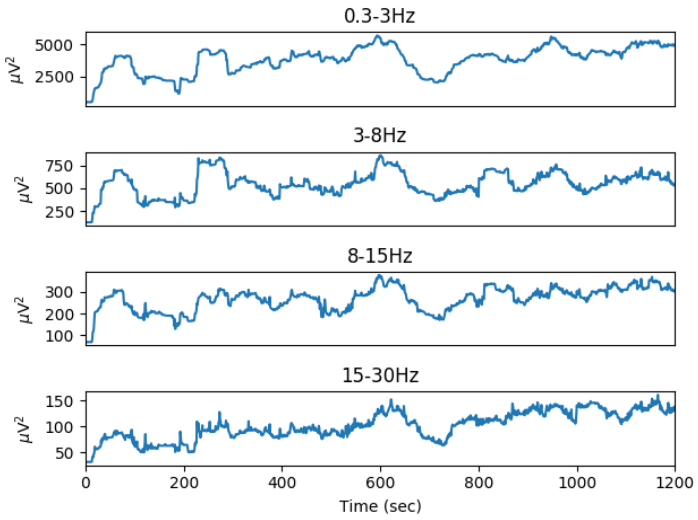


Figure 4.6: 20 minutes of the sub-band power is represented as a median across 8 EEG bipolar channels for each sub-band (27 weeks GA).

Blood pressure

From Figure 4.3 it can be seen that unlike EEG, BP is a slowly evolving signal, where most information is concentrated in the low frequencies (Aletti et al., 2013; Vesoulis et al., 2017). The low frequency range of MAP (from 0.005 Hz to 0.16 Hz) was previously investigated for premature infants (Vesoulis et al., 2017). Therefore, the BP sampling frequency of 1 Hz does not affect the low frequency range of 0 – 0.5 Hz that is of interest here. Figure 4.7 represents the spectrum of the MAP for a preterm neonate (GA=31 weeks).

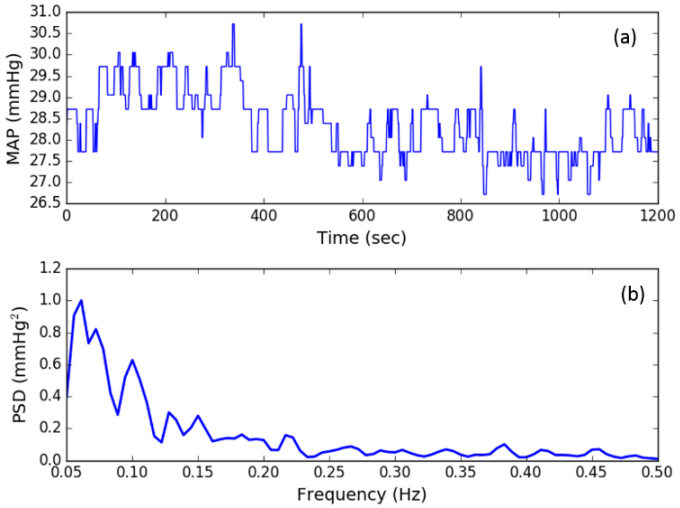


Figure 4.7: An example of the raw MAP trace (a) of a preterm neonate (31 weeks GA) and its spectral power (b). The spectrum was computed using the Welch method on the filtered MAP signal (high pass FIR filter with a cut-off frequency of 0.05 Hz) in order to emphasise the Mayer frequency component at 0.1 Hz. The periodogram was computed for each 180-second long segment of the MAP. The obtained result was then averaged to provide a more accurate estimate of the spectral density. To reduce the effect of the spectral leakage a Parzen window (Andriessen et al., 2003) was applied prior to spectral estimation.

Most of the power in the MAP is concentrated in the low frequencies; it is important to note that a distinctive low frequency component exists at 0.1 Hz. These oscillations with period ~10 seconds were first noted by the German physiologist Siegmund Mayer in the mid-19th century and are known as Mayer waves. The origin of the Mayer frequency component is known to be attributable to the baroreceptor reflex (Julien, 2006). This mechanism helps to maintain BP at a constant level by providing a negative feedback loop. In other words, the increased BP reflexively causes a decrease in the HR, which consequently leads to decreasing BP. At the same time, low BP decreases the activation of baroreflex and causes an increase in HR which allows to restore the normal level of BP.

It has been demonstrated that both EEG and BP have a number of important physiological aspects which should be taken into consideration when working with these signals. An appreciation of the physiology of the premature infant is crucial and will allow for the proper analysis and further interpretation of the obtained results.

4.4 Modelling of interaction between EEG sub-band power and MAP

A general overview of the signal preprocessing, feature extraction and modelling of the interaction between brain activity and BP is shown in Figure 4.8. Measures of linear and nonlinear interaction between EEG and BP are computed. The computed values of coupling are summarized as the median across the entire recording for each newborn (median per subject of 37 hours, IQR = 24 to 48 hours). Per-subject measures of coupling are then contrasted with the corresponding CRIB values. A regression line is fitted using the least squares method. Spearman's rank correlation test (2-tailed) is used to conduct hypothesis tests on the correlation value. A correlation with $p < 0.05$ is considered as statistically significant. In order to provide an insight into the directionality of the interaction between EEG and BP, TE has been computed. The reliability of the obtained results is checked by testing an appropriate null hypothesis for every computed measure of interaction using surrogates. This is done in order to define whether a given empirical non-zero measurement of interaction is statistically different from zero. A general overview of the experimental design is represented in the diagram below (Figure 4.8).

4.4.1 Linear interaction: correlation and coherence

In order to measure the linear association between MAP and brain activity, coherence and Pearson correlation were applied. Unlike the nonparametric Spearman's rank correlation coefficient, which measures the monotonic association between two variables, the Pearson correlation measures a linear coupling. Both correlation and coherence were calculated over a

30-minute moving window with a 30 seconds shift. This window length allows one to focus on the short-term dynamics of both the EEG and MAP signals.

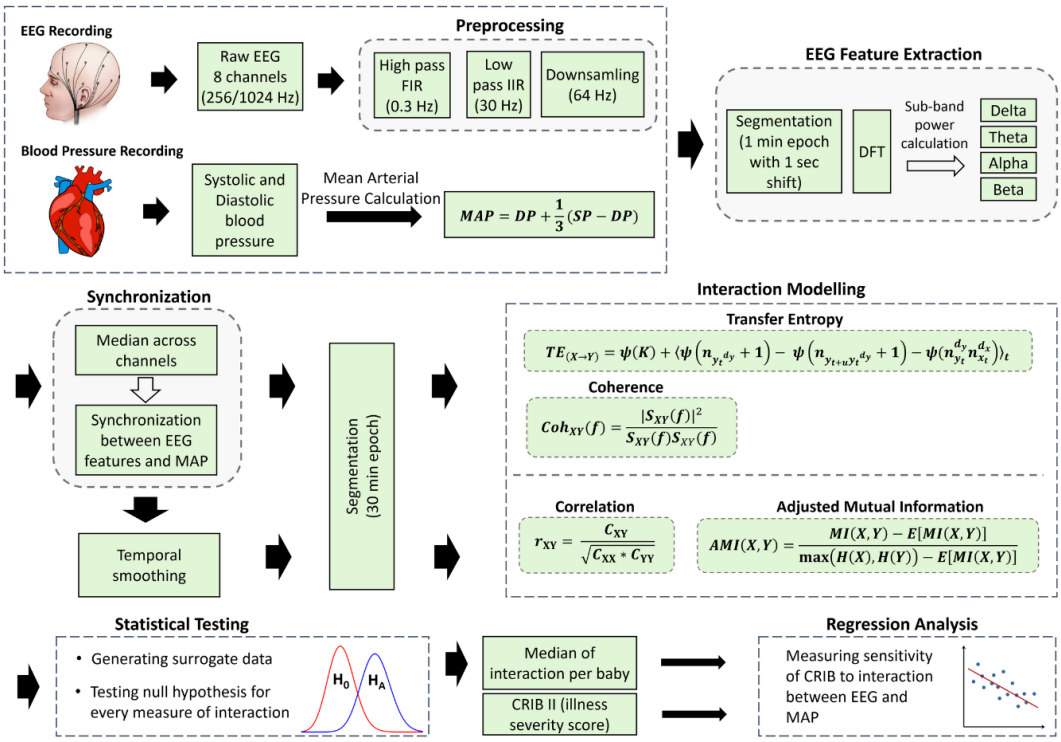


Figure 4.8: Overview of linear and nonlinear modelling of the interaction between EEG and BP signals.

Coherence estimate between BP and four EEG features sampled at 1 Hz was carried out using Welch's method as follows: every 30-minute window is split up into segments of 400 seconds shifted every second. The overlapping segments are then weighted by a Hamming window function. This is then followed by applying DFT to the weighed segments. The individual periodograms are then averaged, which reduces the variance of the individual power measurements. Several different window lengths were tried for the Welch periodogram computation and the segment width of 400 seconds was chosen based on the performance and by visual observation, as a good trade-off between representation of signal periodicity, values of coherence and the presence of distinctive peaks. Every 30-minute window is then represented as a sum of coherence values within a window. An example of coherence computed between EEG and MAP is represented in Figure 4.9.

In order to provide a better insight into the interrelation between brain activity and BP, a cross-correlation was applied. This technique allows flexibility in time and can accommodate the situation in which the change in the feature of one signal did not result in the instantaneous change in the feature of another signal. To investigate whether there is some constant delay between the two signals and to determine whether changes in the EEG proceed changes in the MAP (or vice versa) a cross-correlation was applied. This allowed modelling delays in the BP

and EEG interaction to be introduced at various time lags for each 30-minute window. The cross-correlation was quantified as a maximum value of correlation achieved at a certain time delay. From Figure 4.10 it can be seen that a time lag of the maximum cross-correlation achieved by each epoch varies. This result was consistent for all the preterm neonates. The obtained finding indicates the absence of the constant delay in the association between brain and BP which can be detected using linear methods. This, therefore, may suggest a more complex association between cerebral activity and BP.

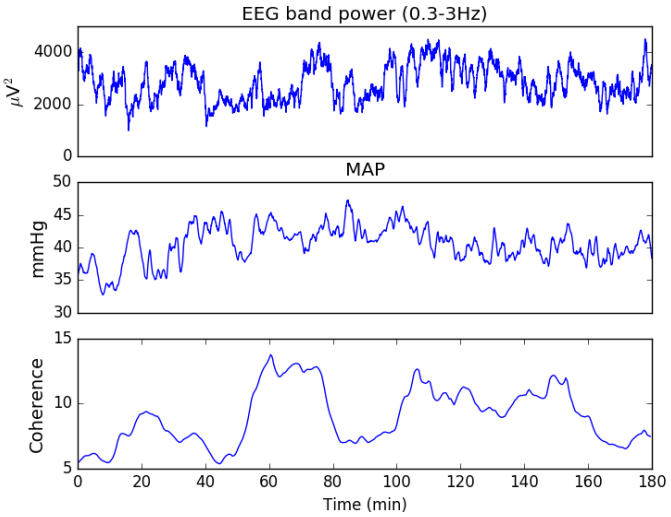


Figure 4.9: Traces of EEG band power and MAP with corresponding coherence between them. Coherence is obtained as a sum of all coherence values within the 30 minutes window (31 weeks GA).

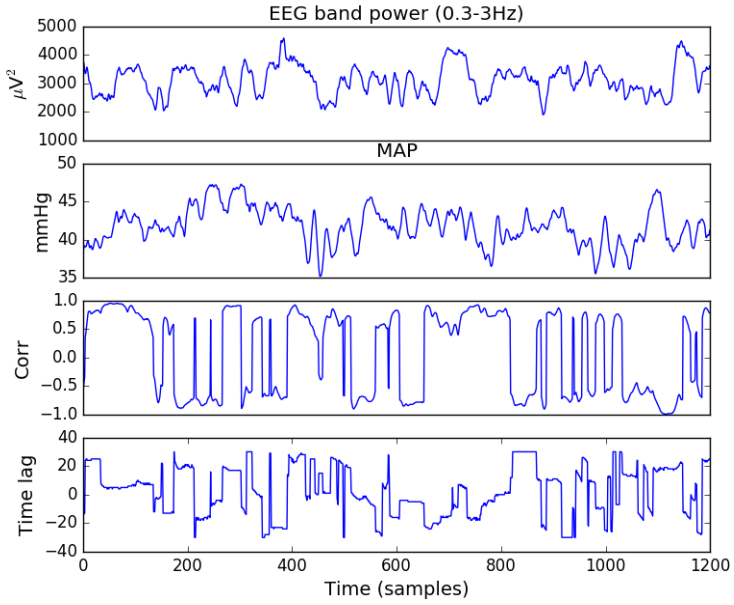


Figure 4.10: An example of a 10-hour trace of the EEG band power (0.3-3Hz) and the MAP of preterm (31 weeks GA) with computed cross-correlation over the 30-minute epoch. The reported cross-correlation is a maximum correlation achieved at a certain time lag for a given epoch. The unit of each time lag here corresponds to a 30-second shift.

4.4.2 Statistical significance

In theory, correlation and coherence between two unrelated sequences is equal to zero. In clinical practice, however, where interaction is empirically measured from a finite number of samples, a non-zero measurement is likely to result even if there is no relationship between the signals. In order to check whether given empirical non-zero measurements of correlation or coherence are statistically different from zero, a null hypothesis of no relationship between signals needs to be tested. The significance of the interaction between the original signals is then estimated against the distribution of interaction values obtained from shuffled surrogates. Surrogate data have been generated by random permutation of original values of MAP and sub-band powers. In other words, after shuffling $\boldsymbol{X}$ and $\boldsymbol{Y}$sequences, instead of $\boldsymbol{p}(\boldsymbol{x}\,|\mathbf{y})$, the surrogate data is distributed as $\boldsymbol{p}(\boldsymbol{x})$. In the present study, 100 shuffled surrogates were generated for every 30 minutes epoch, while leaving the order of epochs unchanged. From Figure 4.11, which represents the PDF of the null hypothesis for correlation computed on all data, we can see that insignificant values range from about -0.4 to 0.4 and therefore, should be ignored. A visual representation of all correlation values and correlation obtained after discarding statistically insignificant values for one preterm is presented in Figure 4.12.

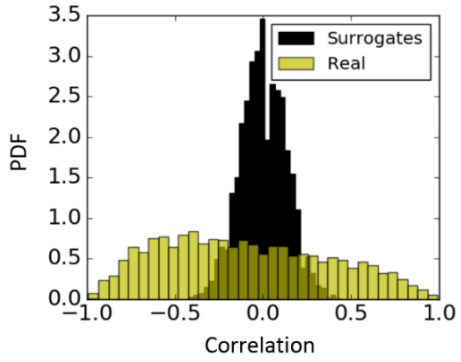


Figure 4.11: Probability density function (PDF) of the Pearson correlation coefficient for surrogates and real data computed on all dataset. Correlation for real data is quantified between MAP and EEG sub-energy (0.3-3 Hz) feature (yellow); correlation between corresponding randomly permuted surrogates of MAP and sub-band energy (0.3-3 Hz) feature (black).

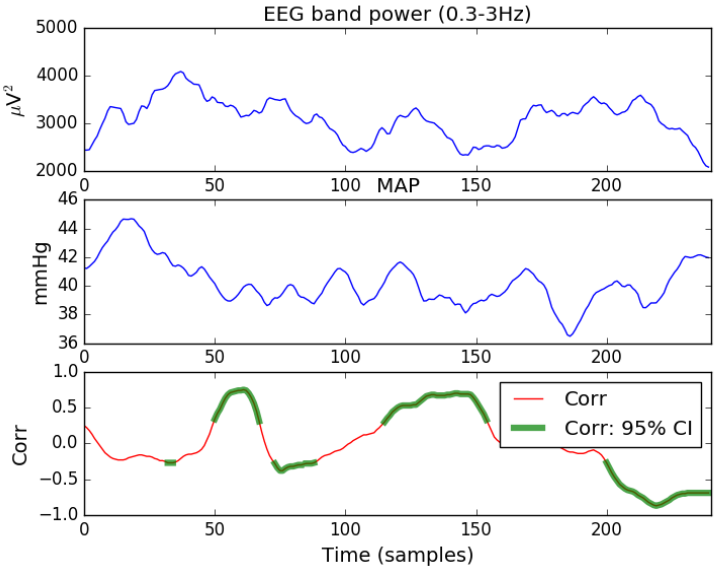


Figure 4.12: All correlation values (including insignificant) (red, thin line) and 95th percentile of the correlation calculated using 100 shuffled surrogates (green, bold line) computed for preterm (31 weeks GA).

4.4.3 Nonlinear interaction: adjusted mutual information

Due to the likely complex relation between brain function and MAP, a nonlinear method of adjusted mutual information (AMI) was applied. AMI is an information theoretic measure of dependency between two random variables. The most common way to calculate AMI from the empirical data is using histogram binning (labelling) in order to estimate the probability density distribution. Figure 4.13 illustrates an example of results of data labelling. It can be seen that a represented 1-hour MAP trace is quantized into 5 labels whereas the 1-hour trace of EEG delta-band powers is converted to 21 different labels.

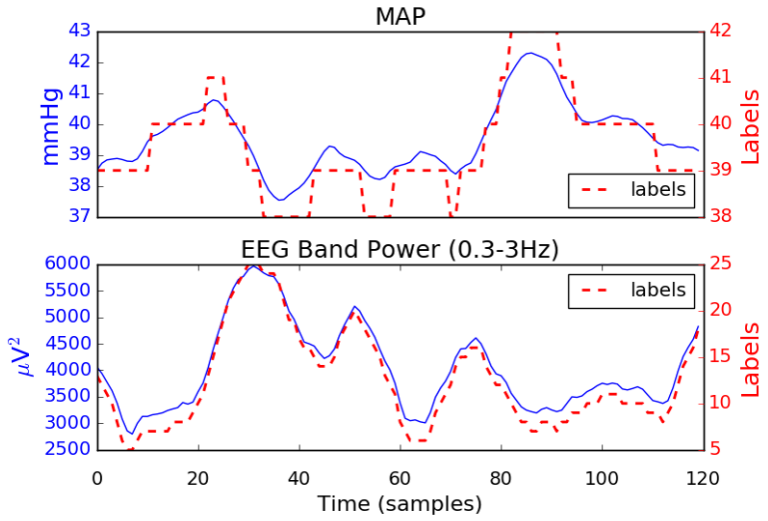


Figure 4.13: One hour of MAP (25 weeks GA) and delta-band energy along with its corresponding labels (dashed line). The shift of a window is 30 sec, which results in 120 values per hour.

As opposed to the conventional MI which increases along with the increasing number of labels, the measure of AMI accounts for the random interaction between sequences. This behaviour can be observed in Figure 4.14, where higher values of MI and lower AMI correspond to a higher number of labels into which the sequences are binned. After reaching some optimal number of labels, the values of AMI start to decrease, penalising the high MI caused by a higher number of labels only. Therefore, binning into an unreasonably high number of labels will result in very low values of AMI implying the absence of shared information between the two sequences beyond that of chance alone. The main advantage of AMI is that for a chosen number of bins, AMI measures the interaction that is adjusted for random chance; the conventional MI measure increases with the increase of random interactions which are caused by the high number of bins.

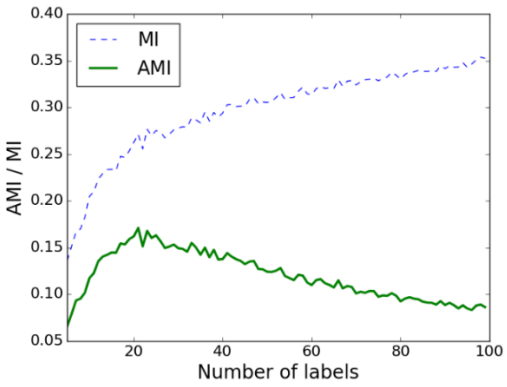


Figure 4.14: The effect of the number of labels on AMI and MI values. Interaction is calculated between the MAP and the EEG power in the 0.3-3 Hz sub-band for one preterm infant (28 weeks GA). Every value is obtained as a mean across all epochs for a given number of labels.

In this study, the MAP signal was binned into a number of labels that is equal to the range of integer MAP values. This number of labels prevents the artificial creation of dynamics in the MAP signal when the MAP fluctuates insignificantly (within ±1 mmHg). The sub-band EEG energy was binned into 35 bins. This number has been chosen as the one that maximizes the values of AMI across all neonates.

MI is a symmetric measure ($MI(X, Y) = MI(Y, X)$) and unlike correlation or coherence, it quantifies both linear and nonlinear dependences. An example is shown in Figure 4.15, where the correlation, AMI and conventional MI are computed between the MAP and the EEG (0.3-3 Hz) band power feature for one newborn. It can be seen that higher levels of both positive and negative correlation result in higher values of AMI and MI, where AMI values are shown to be more conservative as opposed to conventional MI.

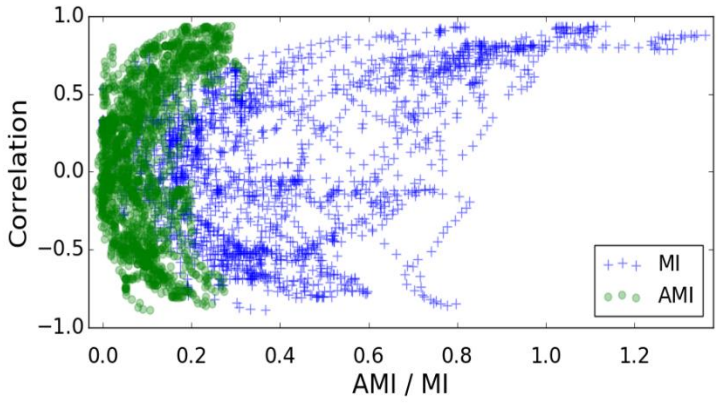


Figure 4.15: A scatter plot of conventional mutual information (MI), adjusted mutual information (AMI) and Pearson correlation. This plot represents measures of interaction between MAP and EEG (0.3-3 Hz) sub-band energy computed for each 30 min window from one preterm (30 weeks GA).

Understanding AMI through simulation studies and surrogates

Conventional MI has already been extensively tested and previously applied to biomedical signals, EEG in particular (Na et al., 2002), (Jeong et al., 2001), while AMI is a relatively new measure (Vinh et al., 2010). In order to check the statistical significance of the measures of interaction based on the MI concept, surrogate tests are conducted in practice (Roulston, 1997; Steuer et al., 2002). The surrogates are obtained by random permutation (shuffling) of the data, which preserves the frequency of the labels but destroys the coupling between the two sequences. The resultant measure of MI is indicative of interaction by chance only. By repeating the random shuffling (100 times in this work), the distribution of the chance-derived MI values is constructed and the statistical test is performed to assess whether the real MI value belongs to the distribution of the MI values obtained by chance. As the AMI measure explicitly accounts for chance, i.e. increasing the number of labels (bins) will not increase the value of AMI, therefore, there is no need to check the statistical significance of the obtained results with respect to chance. In order to illustrate this characteristic of AMI, artificial data were generated, which allows for the control of the different parameters, such as the level of noise and correlation between data while measuring the level of coupling. Two artificial sequences were simulated as a sum of sinusoids with and without random noise added (Figure 4.16). The surrogate test was performed on these sequences and the obtained result is presented in Figure 4.17.

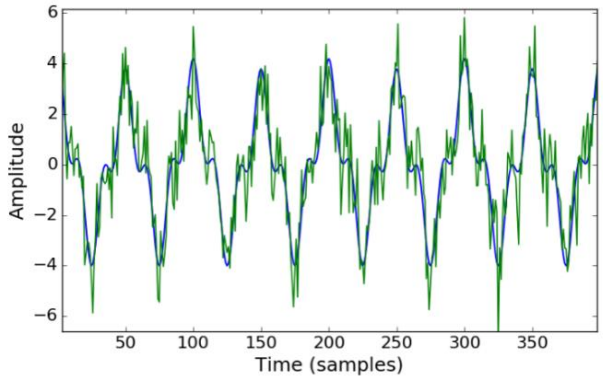


Figure 4.16: Trace of artificially generated toy data. $y_0 = 2.8 * sin(2\pi * 1000) + 1.2 * sin(2\pi * 3000x) + 0.2 * sin(2\pi * 500x)$ in blue and $y = y_0 + N(0,1)$ in green. Pearson correlation coefficient between y_0 and y is equal to 0.91.

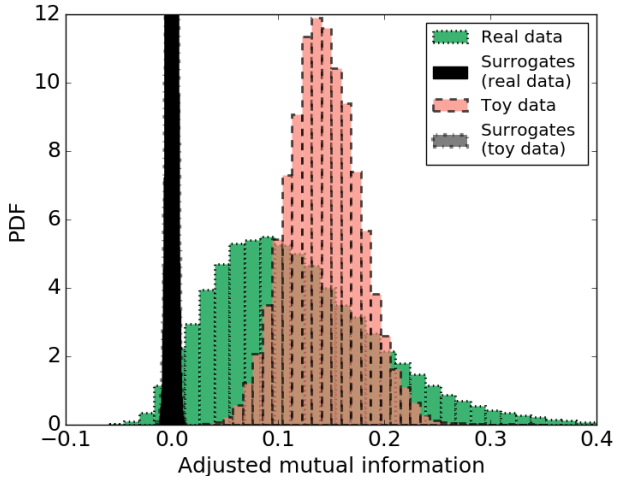


Figure 4.17: The probability density function of adjusted mutual information obtained from all data (25 subjects). AMI for real data (MAP and sub-band energy) (green) and its permuted surrogates (black); toy data (artificially created correlated signals) (red) and its permuted surrogates (grey). The distributions of the surrogates are overlapped and clipped.

It can be seen from the histograms that the level of coupling measured by AMI for sequences with random interaction (surrogates) is centred on zero whereas the AMI values for the correlated data (red) is centred on 0.15. This confirms that AMI accounts for the chance and that non-zero AMI values measure the inherent level of interaction in the two sequences. Figure 4.17 also shows the same plot for real data from the database. For every 30-minute window, AMI is calculated for the original sequence of MAP and EEG feature values and for permuted sequences. The distributions of the AMI values for both the real data and the toy data are clearly separated from corresponding surrogates, which implies the reliability of the calculated measure and indicates the presence of non-random interactions in the real data.

The AMI values are quite conservative, even for strongly correlated data. In order to check the sensitivity of the AMI measure to random noise and establish an intuitive connection with

correlation, different noise levels were added to the artificially generated toy data as follows: $y = y_0 + N(0, n)$, where n is the standard deviation of random samples drawn from a Gaussian distribution (Figure 4.18).

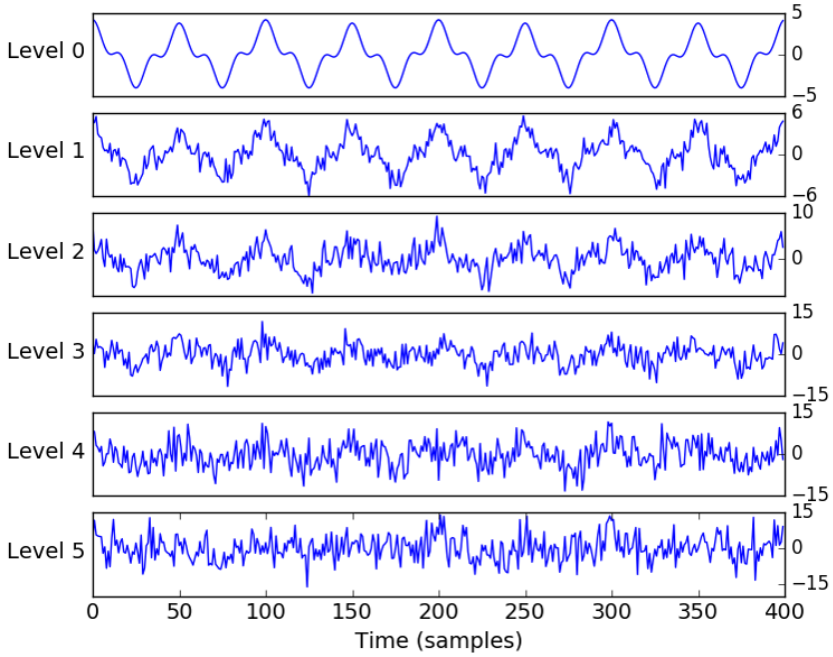


Figure 4.18: Artificially generated data with different levels of noise added.

A zero coupling baseline was set by measuring the interaction between two Gaussian independent and identically distributed (IID) processes. It can be seen from Figure 4.19, that the coupling of the two IID processes is centred around zero. This result shows an absence of interaction between two random sequences as measured by both AMI and correlation.

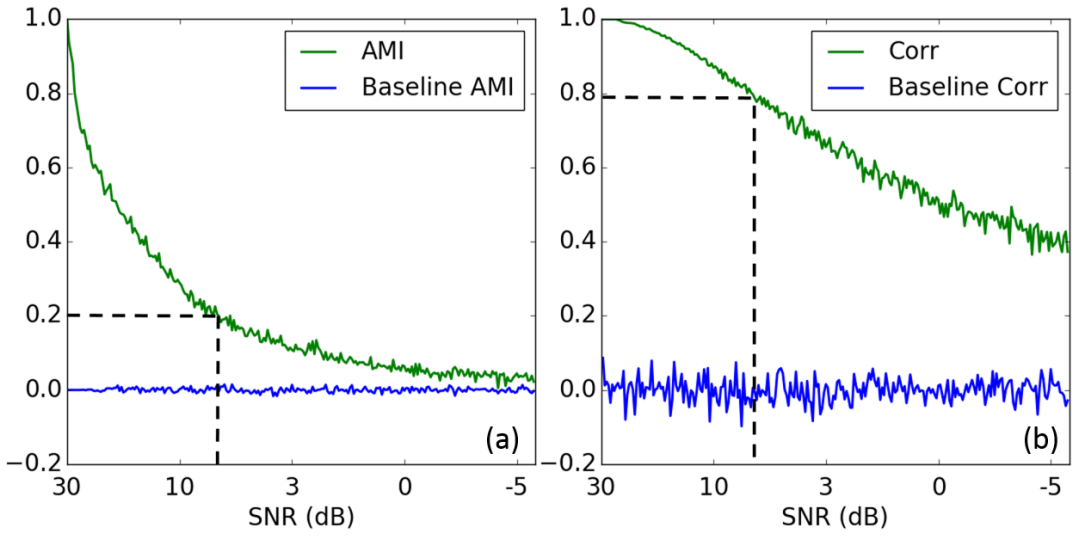


Figure 4.19: Effect of noise on AMI (a) and correlation (b). Baselines of zero coupling (blue, zero centred) for both measures are represented as Pearson correlation and AMI between two Gaussian independent and identically distributed processes.

We can also observe that high levels of noise have a greater impact on AMI than correlation, where an AMI of 0.2 corresponds to a correlation coefficient of 0.8. Comparing Figure 4.19 and Figure 4.17, it also can be seen that the operating range of AMI values for real physiological data from 0.05 to 0.25 (Figure 4.17) corresponds to Pearson correlation coefficients of about 0.5-0.8 (Figure 4.19, dashed line). This justifies rather low and conservative values of AMI obtained even for clearly correlated data.

4.4.4 Directionality of interaction

MI does not contain any directional information as it is a symmetric measure, where $MI(X,Y) = MI(Y,X)$, and therefore it is not effective at predicting future events from the data or deriving the causality between two sequences. As described in Chapter 3, TE is an extension of MI which takes into account the direction of informational flow.

In this work, the number of nearest neighbours was chosen to be $K = 4$, as recommended in (Kraskov et al., 2004) to balance bias (which decreases for larger K) and variance (which tends to increase for larger K). In order to find an embedding dimensions for MAP and the EEG features, we have used the measure of active information storage (AIS), which defines the past information of the process that can be used to predict its future (Lizier et al., 2012). AIS A_X for the sequence X is defined as the expected MI between the past state of the process X_t^d (as $d \rightarrow \infty$) and its next state X_{t+1}:

$$A_X(d) = MI(X_t^d, X_{t+1}) \tag{4.1}$$

Here d is the embedding dimension, which captures the underlying state of the process X for a Markov process of order d.

The proper embedding history for signals can be set according to the peak values of AIS plotted against the embedding history (Lizier, 2014). Results represented in Figure 4.20, indicate that embedding history of $d = 3$ is sufficient for BP, while for all four EEG sub-band powers embedding dimension is set to $d = 2$. TE was estimated using an open source toolbox JIDT (Lizier, 2014).

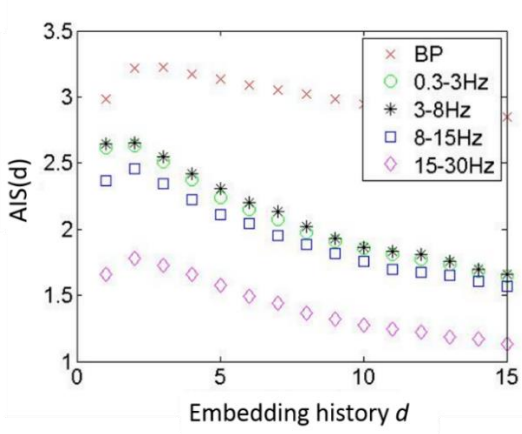


Figure 4.20: Active information storage for BP and four EEG sub-bands.

When conducting statistical analysis for TE results, it is necessary to take into consideration the chosen values of the embedding dimensions e.g. the length of history we are checking in Y when trying to predict X. The reliability of $TE_{(Y\rightarrow \mathrm{X})}$, is tested against $TE_{(Y^s\rightarrow \mathrm{X})}$, where Y^s is a permuted surrogate, created by shuffling vectors $y_t^{d_y}$ (Lindner et al., 2011; Lizier, 2014),(Vicente et al., 2011). As a result, the obtained surrogates preserve $p\left(x_{t+1}\middle|x_t^{d_x}\right)$, but not $p(x_{t+1}|x_t^{d_x}, y_t^{d_y})$.

4.5 Results of the association of coupling measures with the illness severity score

The results of the association between CRIB scores and the computed measures of coupling are presented in Table 4.2. Insignificant values of correlation were defined using the 95% CI obtained using the bootstrap method (random sampling with replacement).

Table 4.2: Correlation coefficient and 95% CI (in brackets) between CRIB score and coupling measures (correlation, coherence, AMI, and TE) between MAP and four EEG sub-band powers.

	(EEG 0.3-3 Hz & MAP) vs CRIB	**(EEG 3-8 Hz & MAP) vs CRIB**	**(EEG 8-15 Hz & MAP) vs CRIB**	**(EEG 15-30 Hz & MAP) vs CRIB**
Correlation	NS	NS	NS	NS
Coherence	NS	NS	NS	NS
AMI	r: -0.57 (-0.78, -0.22)	NS	r: -0.42 (-0.68, -0.03)	NS
TE (EEG to MAP)	r: 0.428 (0.11, 0.68)	r: 0.44 (0.21, 0.66)	r: 0.416 (0.13, 0.67)	NS
TE (MAP to EEG)	NS	NS	NS	r: -0.436 (-0.68, -0.17)

Correlation coefficients were determined to be significant when their 95% CI excludes zero.

No statistically significant association was found between the CRIB scores and the level of linear coupling between MAP and all four EEG sub-band energies measured using both

correlation and coherence. This result is obtained after discarding statistically insignificant relationship between EEG sub-bands and MAP and replacing them with zero.

Figure 4.21 shows the association between the AMI (using MAP and the sub-band energy 0.3-3 Hz) and the CRIB scores. As expected, a low MAP correlates with high CRIB scores, where higher risks of mortality are associated with lower MAP values (r=-0.503, p=0.01). This association could be indirectly associated to gestational age, as CRIB scores and MAP are both dependent on the GA. At the same time, the higher CRIB scores were not correlated with changes in any of the EEG energy bands (0.3-3 Hz: r=0.24, p=0.2; 3-8 Hz: r=0.08, p=0.7; 8-15 Hz: r=-0.12, p=0.6 and 15-30 Hz: r=-0.2, p=0.3). It can also be seen from Figure 4.21 that the CRIB score has marginally higher correlation with the developed measure of interaction between signal dynamics, AMI, than with MAP (r=-0.57, p=0.003 vs r=-0.503, p=0.01). A statistical test of the equality of the two correlation coefficients obtained from the same sample, with the two correlations sharing one variable (CRIB) did not show significant difference between them. The level of correlation between the CRIB score and the AMI for MAP with other EEG sub-band energies was significant only for 8-15 Hz sub-band (3-8 Hz: r=-0.3, p=0.13; 8-15 Hz: r=-0.42, p=0.04 and 15-30 Hz: r=-0.36, p=0.08). There is also a statistically significant correlation between AMI and MAP for two sub-bands only (0.3-3 Hz: r=0.41, p=0.04; 3-8 Hz: r=0.35, p=0.09; 8-15 Hz: r=0.38, p=0.06; 15-30 Hz: r=0.54, p=0.004). Increase in MAP was associated with increasing GA (p=0.001, r=0.6) and an increase in EEG (15-30 Hz) spectral power (r=0.541, p=0.005).

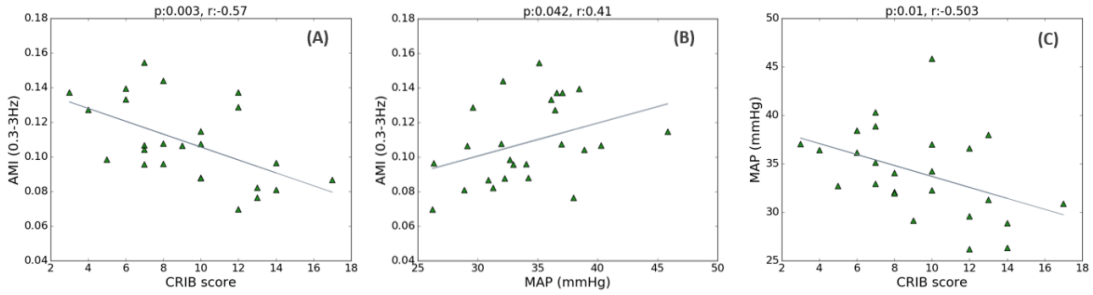


Figure 4.21: The relationship between CRIB, MAP and EEG energy. (A) CRIB score and AMI between MAP and EEG energy (0.3-3Hz); (B) MAP and AMI; (C) CRIB score and MAP. Every point on the scatter plots represents 1 newborn.

If only the data recorded during the first 24 hours of life of the preterm is considered (Figure 4.22), the association between AMI for the sub-band energy 0.3-3 Hz and the CRIB score improves with respect to the values computed over the whole recordings (r=-0.57, p=0.003 vs r=-0.65, p=0.001). At the same time, the association between AMI and MAP becomes insignificant (r=0.416, p=0.054).

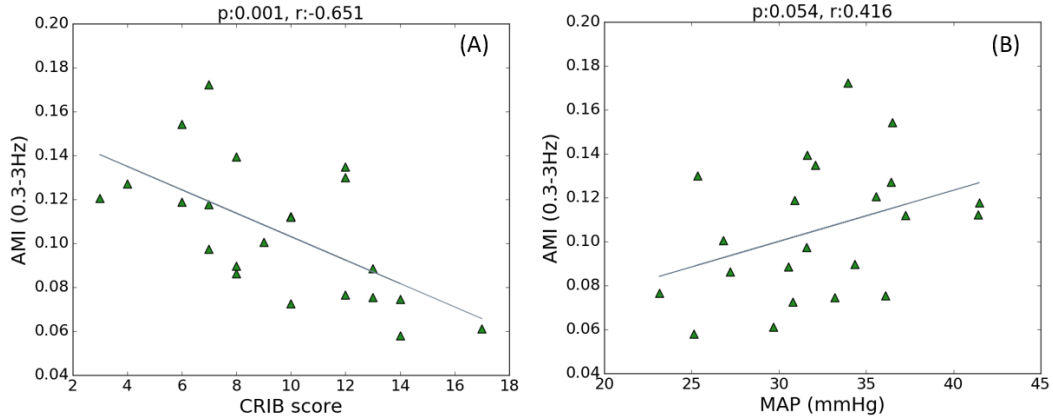


Figure 4.22: The relationship between CRIB, MAP and EEG energy for the data during the first 24 hours of life only.

While AMI is focused on the detection of the significant coupling between sequences, TE also detects the direction of coupling. As represented in Figure 4.23, no association was found between TE (MAP to EEG (0.3-3 Hz)) and CRIB score (r=-0.089, p=0.672), however the transfer of information in the opposite direction, from EEG (0.3-3 Hz) to MAP, showed an association with CRIB scores (r=0.428, p=0.033). Results of TE for the other three EEG sub-band powers are represented in Table 4.2. From Figure 4.24 it can be seen that TE of real data is separated from its corresponding surrogates. This indicates the reliability of the obtained TE values. At the same time, TE from MAP to EEG is lower than corresponding values of TE from EEG to MAP.

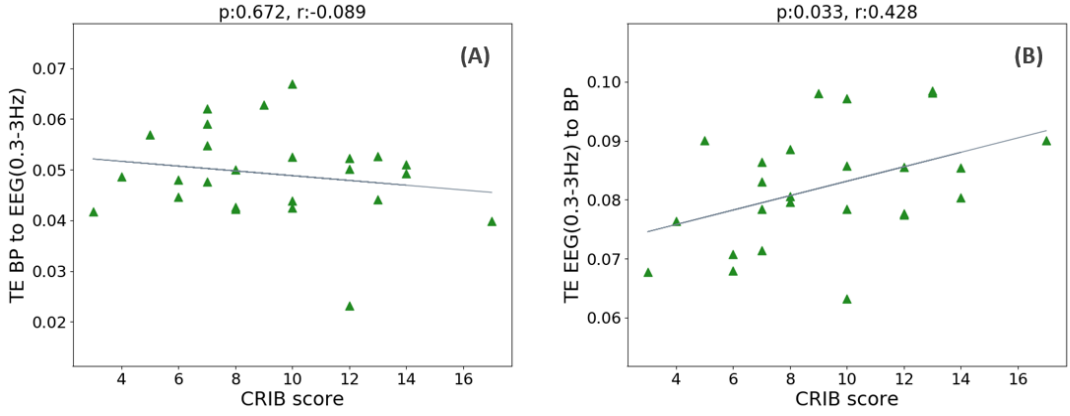


Figure 4.23: The relationship between the CRIB score and interaction between MAP and EEG (0.3-3 Hz) energy. (A) CRIB score vs TE from MAP to EEG (0.3-3 Hz); (B) CRIB score vs TE from EEG (0.3-3 Hz) to MAP.

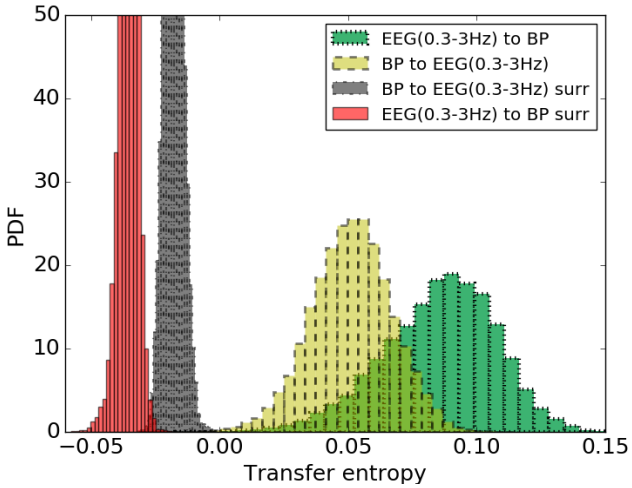


Figure 4.24: The distribution of TE values for real data and randomly permuted surrogates. TE is quantified from MAP to EEG (0.3-3 Hz) for real data (yellow) and corresponding randomly permuted surrogates (black). TE from EEG (0.3-3Hz) to MAP for real data (green) and its randomly permuted surrogates (red).

4.6 Decision support tool

The results in this chapter indicate that the physiological reaction to the changes in BP is associated with lower risks to the preterm. Hypotension in preterm neonates still has no clear definition and the decision on whether it should be treated remains challenging, which has resulted in the variability of treatment procedures. The finding obtained in this study can potentially contribute towards the generation of a hypothesis in the field of hypotension management and the interrelation between cerebral activity and BP for preterm neonates.

An example of the visualization of interaction between the MAP and cortical activity quantified by the EEG energy in 0.3-3 Hz (delta) sub-band is represented in Figure 4.25. It can be seen that starting from the fourth hour of recording the level of MAP drops and stays below the GA threshold (in weeks). In the most common clinical setting, where clinicians follow the GA-based rule for hypotension management, this level of MAP would be considered as abnormally low and most likely treatment would be initiated. At the same time from Figure 4.25 it is clear that for a given hypotensive region the level of interaction between the brain and BP represented with the cumulative average of AMI is around 0.12. According to the previously reported result of the interaction between brain activity and BP, and its association with the illness severity score (Figure 4.21), the AMI value of about 0.12 corresponds to comparatively low levels of risks for the preterm. Therefore, this information might suggest that following a so-called ‘permissible approach’ for hypotension management, which excluded any interventions, may be appropriate.

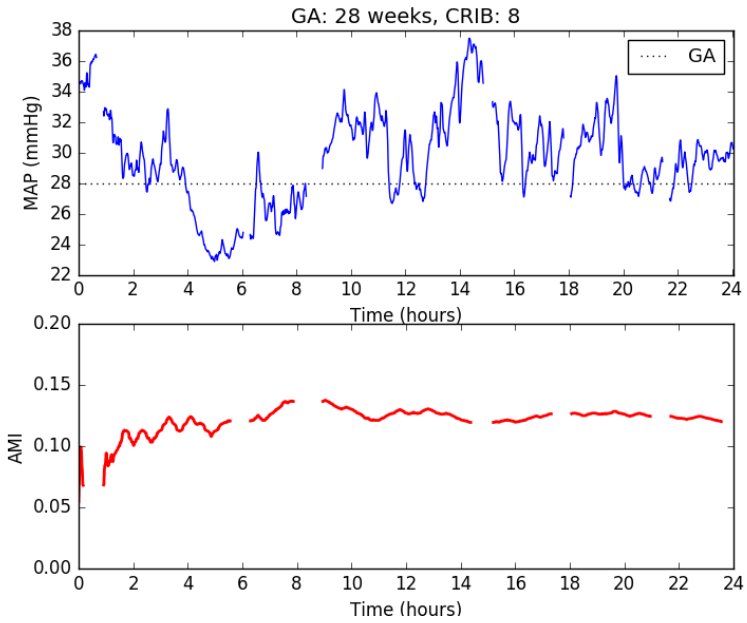


Figure 4.25: Visualisation of the MAP along with the interaction of the MAP and EEG power (0.3-3 Hz) (represented as a cumulative average of AMI) for the preterm neonate (28 weeks GA). Missing values on the traces correspond to the eliminated regions affected by artefacts.

Two channels of EEG are now routinely used in infants who are suspected of having a brain injury. Therefore, we anticipate that a module in a bedside monitor incorporating the algorithm to compute and visualise the measure of interaction between BP and cerebral activity would be feasible. It will provide real-time decision support for more efficient management of hypotension in preterm neonates. Therefore, instead of relying solely on the GA-based threshold rule, the level of interaction between two systems will help to guide clinicians to provide a patient-specific treatment.

4.7 Discussion

4.7.1 Nonlinear relationship between the MAP and EEG sub-band powers

The main finding of the present study is related to nonlinear measures of interaction between cerebral activity and MAP for preterm neonates. A statistically significant association of the CRIB scores with AMI is observed for the low frequency (0.3-3 Hz) sub-band energy of the EEG. Maturational features for neonatal EEG vary across GAs. Most of the preterm EEG power is known to be concentrated in the lower frequencies. Delta (0-3.5 Hz) activity is a major characteristic of the preterm EEG that evolves as the infant matures and disappears between 38 and 42 weeks of gestation (Pavlidis et al., 2017).

The obtained correlation of CRIB with AMI is higher than that of CRIB with MAP (r=-0.57, p=0.003 vs r=-0.503, p=0.01), although the difference was not statistically significant.

However, it is worth emphasising that AMI is independent of the absolute values of both MAP and EEG energy and measures only the coupling between signal dynamics. Thus, the measure of dynamic interaction correlates with the median MAP per baby. We found lower levels of coupling at lower values of MAP. This indicates that low BP affects the interaction measured by AMI as do poor CRIB scores. Additionally, the results of simulation and surrogate tests which were used to detect random coupling, support the reliability of the obtained AMI values.

From Figure 4.2 it can be seen that every preterm neonate has different duration and timing of the recordings. When considering EEG and BP data during the first 24 hours after birth, an association between AMI (0.3-3 Hz) and CRIB score has improved (r=-0.57, p=0.003 vs r=-0.651, p=0.001). These results indicate that the coupling between cerebral activity and BP is more sensitive to the risk index of the preterm during the first hours of life.

4.7.2 Linear relationship between MAP and EEG sub-band powers

Linear measures of interaction such as correlation and coherence have been previously applied to adult physiological signals (Pfurtscheller et al., 2012) including NIRS, EEG, ECG, and BP, where the analysis has been conducted on preselected 5-minute epochs from 19 subjects. In our work after discarding insignificant correlation coefficients (Figure 4.11) using surrogate tests, both correlation and coherence measures have indicated that these linear measures of interaction between brain activity and MAP failed to find an association with the underlying illness severity scores of the preterm infant.

Brain function is known to be a complex system of nonlinear processes and therefore it is likely that nonlinear methods would be more appropriate when measuring the interaction between EEG and MAP. The results of this study showed that a nonlinear method of coupling between MAP and EEG features measured using AMI is more sensitive to noise compared with the linear correlation method (Figure 4.19). According to (Netoff et al., 2006) nonlinear measures are indeed very sensitive to noise and linear methods sometimes present better properties in this sense (Pereda et al., 2005). However, both linear and nonlinear approaches assess different aspects of the interdependence between the signals and provide a more comprehensive picture of the analysed data. Therefore, it is a good practice to use both methods to ensure that all the information available from the signals has been obtained and properly analysed with statistical and reliability tests (surrogates).

4.7.3 Directionality of interaction

The detection of the presence of a dominant direction for the coupling between physiological systems can also provide an insight into their mutual interdependency. TE was previously used for establishing a directed information structure between brain regions (Lizier et al., 2011). In

the area of cardiovascular physiology, this technique was utilized to define causal relationships that explain sources of variability in the regulation of cerebral hemodynamics (Katura et al., 2006). In this study, the TE measure is used to provide an indication of the causal relationship of processes that occur between brain activity and MAP with respect to the illness risk of the preterm infant. The TE results passed the surrogate test which indicates that the directionality in the MAP and EEG sequences is present beyond the level of chance. The higher values of TE from EEG to MAP in comparison with the opposite direction indicate greater information transfer from EEG to MAP. The strength of this directionality apart from being greater also correlates with the CRIB scores for the first three EEG sub-band powers as shown in Table 4.2. This may indicate that sicker preterm infants have a higher level of information flow from brain activity to MAP for a wide range of frequencies (lower than 15 Hz).

The neuronal activation followed by hemodynamic changes has been previously reported in (Steinbrink et al., 2006), (Mangia et al., 2009). At the same time Caicedo et al., (2016) and Roche-Labarbe et al., (2007) have shown that changes in cerebral oxygenation assessed by NIRS are likely to precede changes in EEG; in (Caicedo et al., 2016), the causality is represented by higher TE values from NIRS to EEG. Unlike the previous study where the EEG was recorded using only the C3-C4 channel, in this study eight bipolar EEG channels were incorporated, which enables better coverage of the preterm brain (Pavlidis et al., 2017). Additionally, in our study, the EEG was analysed through four sub-band energies from long duration unedited signals, whereas a single root mean square measure of EEG energy was used in (Caicedo et al., 2016). Moreover, the interaction between EEG and MAP was explored in the context of CRIB scores, whereas in (Caicedo et al., 2016) the interaction was measured between EEG and NIRS under a sedation protocol. These differences make it difficult for a direct comparison of the results. The physiological mechanism of autoregulation for premature babies is not fully understood and to the best of our knowledge, this is the first study to investigate the interaction and the directional information flow between MAP and EEG for preterm neonates. Therefore, in order to have a better understanding of the physiological mechanism of connectivity between brain and MAP, further studies are warranted.

4.7.4 General discussion

The results obtained in this study allow us to hypothesise that the stronger coupling between brain activity and MAP, as quantified by AMI, is related to the physiological status of a preterm (Figure 4.21). At the same time, stronger directionality in the interaction is associated with an increased risk of mortality (Figure 4.23). This hypothesis is schematically represented in Figure 4.26 based on the cerebral autoregulation curve. Here, the functioning cerebral autoregulation plateau which represents normal physiological wellbeing corresponds to a

stronger interaction between EEG and MAP (higher AMI values) and weaker EEG-to-MAP directionality (TE values). If it is argued that the overall status of the infant is affected by hypotensive periods, then the problem becomes one of comparing the MAP with some lower threshold, such as MAP=GA. However, there is not one common threshold for every infant. In this work, it is proposed for a particular infant, that when the MAP falls below this unknown threshold that the dynamic (rather than static) interaction changes. Therefore, identifying the change in slope of the autoregulation curve, through the proposed measures of dynamic interaction, can be used as a proxy for identifying the threshold.

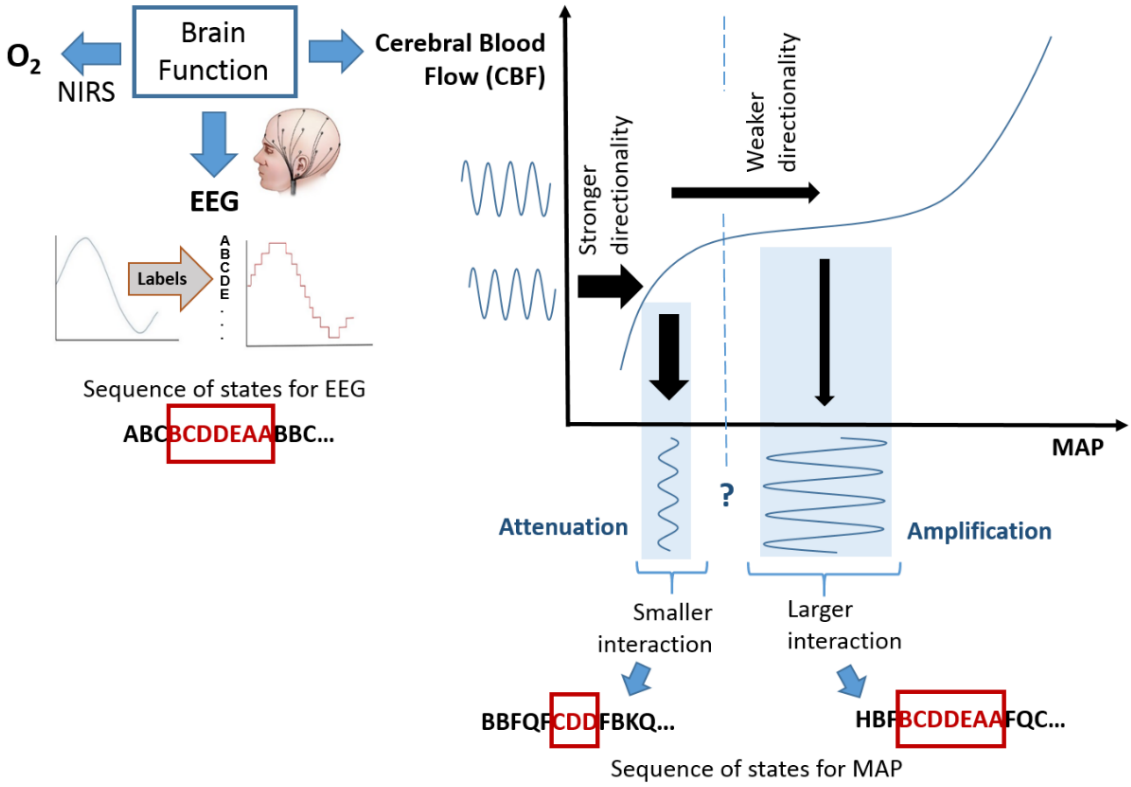


Figure 4.26: Schematic representation of the coupling between EEG and MAP using the autoregulation curve. The plateau of the autoregulation curve is used as a benchmark of normal brain function. The MAP and EEG are represented by the sequence of states denoted with letters. A higher level of interaction is implied by a longer overlap in the sequences (same patterns). When the MAP falls below an unknown threshold, the dynamic interaction between EEG and MAP changes. A higher risk of mortality which is represented with higher CRIB scores is shown to be associated with a smaller interaction between EEG and MAP and a stronger directionality of this interaction (from EEG to BP).

In this study, we hypothesise that during periods of low BP, EEG activity would change, which may lead to a change in the coupling between BP and brain activity and therefore may be indicative of newborn wellbeing. The association between BP and EEG activity in neonates is complex and can be influenced by the blood supply to the brain as well as autoregulatory mechanisms. It was previously reported that cerebral autoregulation is poorly developed in the preterm and is influenced by many factors (Jayasinghe et al., 2003). This is supported by a

study (Victor et al., 2006a), where forty preterm infants were assessed, but no statistically significant relationship between BP and cerebral electrical activity was identified. Similar to (D. Shah et al., 2013; Victor et al., 2006a) we observed an increase in MAP with increasing GA (r=0.614, p=0.001) (Figure 4.27).

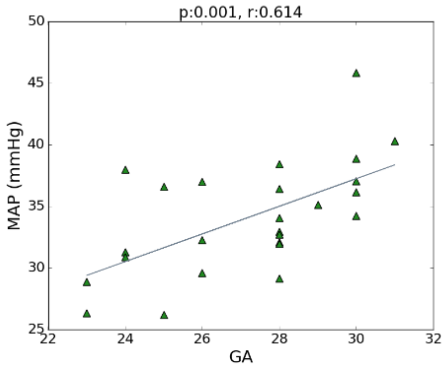


Figure 4.27: An association between MAP and GA.

At the same time, an increase in EEG (15-30 Hz) spectral power was also associated *(r*=0.54, *p*=0.005*)* with increasing MAP (Figure 4.28).

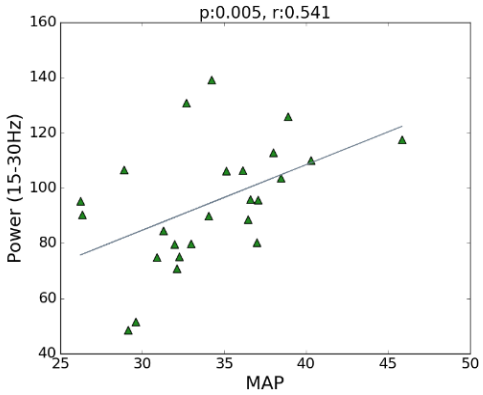


Figure 4.28: An association between EEG power (15-30 Hz) and MAP.

No changes in absolute spectral powers were associated with the GA. In contrast, other studies reported changes in absolute and relative spectral powers with increased postmenstrual age (maturation) (Gagliardi et al., 2004; Niemarkt et al., 2011; Okumura et al., 2006); there may be a number of reasons for this. It is known that relative power, which is not considered in the current study, is better at capturing the maturational changes in the preterm brain than absolute power, as this measure is more sensitive to the changes in EEG discontinuity. Another reason could be related to different EEG processing routines and to procedures which were used to select EEG segments. For instance, in (Bell et al., 1991) spectral analysis was performed on only eight 4-second epochs from each recording. The study in (Okumura et al., 2006) has also analysed preselected short EEG epochs with a duration of only 10 seconds. In contrast, in our

study, spectral analysis was performed on the continuous multi-channel unedited EEG recordings from 25 subjects with a total duration of 957 hours.

Holm-Bonferroni correction for the multiple comparison (n=5) showed that in order to be significant at alpha level of 0.05, the first-ranked (smallest) *p* value needs to be smaller than 0.01. After the correction is applied, the association of AMI between MAP and EEG 0.3–3 Hz and CRIB score remained statistically significant which is the main result of the study (Table 4.2). This study, however, is of exploratory nature and aims at open-ended hypothesis generation. Several researchers have recently argued that *p* values lose their meaning in exploratory analyses due to unknown inflation of the alpha level (Nosek and Lakens, 2014; Wagenmakers et al., 2012). This allows *p* values to serve as a guide for the hypothesis to be tested in further confirmatory research.

4.7.5 Limitations

Preterm cortical activity can be characterised by a number of maturation features (Pavlidis et al., 2017), such as continuity, sleep states, and others. In this study, we assessed the first days of life of the preterm only, where every infant was represented by a single summary measure (median across the recording). As a result, this did not allow us to investigate the possible impact of the cyclical activity, such as sleep states, on the coupling across time.

In order to better represent the population of preterm neonates, further confirmatory research should be conducted on a larger cohort of preterm neonates with a wider range of CRIB scores. This, however, may be a challenging task, as continuous multichannel long EEG recordings are very difficult to obtain for the population of preterm neonates. These babies are extremely vulnerable and any intervention should be agreed with the neonatologist. At the same time, neonates with high CRIB score are very sick and it is difficult to get permission for such an intervention.

4.8 Conclusions

In this chapter, we have investigated the relationship between short-term dynamics in BP and EEG energy in the preterm on a large dataset of continuous multi-channel unedited EEG recordings. The coupling between EEG and BP is computed using both linear and nonlinear measures for 25 preterms. Our findings suggest that nonlinear measures of interaction are more suitable when measuring coupling between the complex system of brain function and BP. The results are tested with surrogate reliability tests and contrasted with the preterm wellbeing represented by the CRIB score. The findings reported in this study have indicated that a higher risk of mortality for the preterm is associated with a lower level of nonlinear interaction

between EEG and MAP which is measured by AMI. The computation of the proposed measure of interaction is independent of absolute values of MAP and GA-based thresholds. It has been shown that higher CRIB scores are also associated with higher levels of information flow from EEG-to-MAP as measured by TE. This allows us to hypothesise that the normal wellbeing of a preterm neonate can be characterised by a strong nonlinear coupling between brain activity and MAP, whereas the presence of weak coupling with distinctive directionality of information flow may be associated with an increased risk of illness severity in preterms. Obtained findings can potentially contribute to the more efficient treatment of hypotension in preterms. The visualiation of the interaction between brain activity and BP can serve as an additional source of information for patient-specific treatment procedure.

Chapter 5: Prediction of short-term health outcome using multimodal physiological signal analysis and boosted decision trees

Both EEG and ECG are extensively used for assessing aspects of newborn health. In this chapter, the relationship between multiple clinical modalities is investigated in order to predict the short-term health outcome in preterm infants. In particular, the predictive capability of HRV and EEG for the estimation of short-term health outcome is assessed and the predictive power of both HRV and EEG features during episodes of low BP is studied. A decision support tool for the continuous estimation of the probability of neonatal morbidity based on the observed physiological data and the boosted decision tree classifier is proposed.

5.1 Introduction

Due to the absence of any firm guideline on the assessment of hypotension in preterm neonates, it is important that not only BP is monitored but also other physiological signals including EEG and ECG. The information extracted from both EEG and ECG recordings is extensively used for assessing various aspects of newborn health. Promising results have been obtained for automated computer-based outcome prediction in full-term neonates using a combination of multimodal features including HRV and EEG (Temko et al., 2015). For term neonates, a significant association between HRV, the severity of hypoxic ischemic brain injury and long-term neurodevelopmental outcome at two years of age was reported for 61 full-term neonates (Goulding et al., 2015).

Both EEG and ECG have defined ranges which correspond to the normal well-being of the neonate. This information allows for the proper interpretation of the current health status as well as the potential to predict the future well-being of the neonate. While several studies have

identified an association between HRV, EEG and neonatal health outcomes in the term and preterm infants, there is still a lack of understanding about this relationship in the context of low BP episodes in preterm infants. This study investigates the relationship between these three different modalities – EEG, ECG, BP – and the short-term health outcome in preterms with a GA less than 32 weeks. In particular, the predictive capacity of HRV and EEG for the estimation of short-term health outcome is assessed and the predictive power of both HRV and EEG features during the episodes of low BP is studied.

5.2 Dataset

This study concentrates on the short-term health outcome of the preterm neonate quantified by the CCS, which is assigned at discharge from the NICU (Figure 5.1). The score summarises the presence of at least one of the five major neonatal complications. A more detailed explanation is provided in Chapter 2. The study had full ethical approval from the Clinical Research Ethics Committee of the Cork Teaching Hospitals.

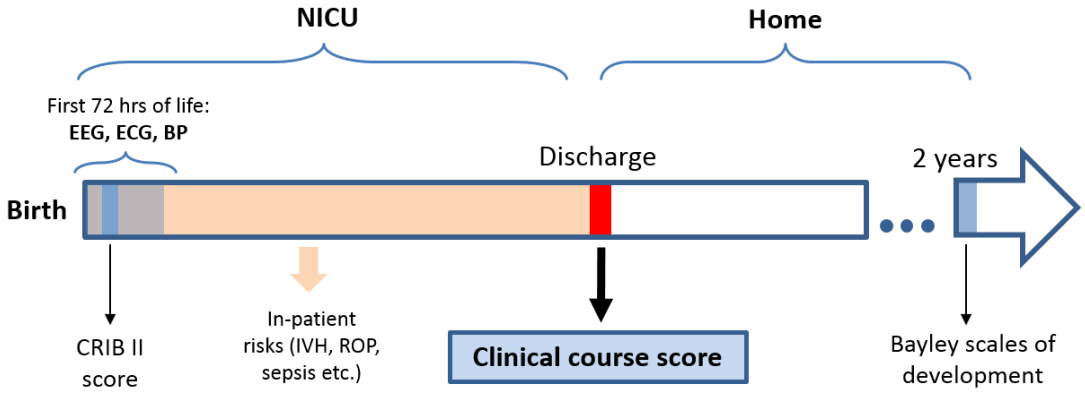


Figure 5.1: The timing of illness scores assigned to an infant during the course in the neonatal intensive care unit (NICU) through to the neurodevelopmental follow-up at 2 years of age. In-patient risks include major neonatal complications: IVH (intraventricular haemorrhage), cystic periventricular leukomalacia, necrotizing enterocolitis, infection (sepsis), retinopathy of prematurity (ROP). The diagram is adapted from Lloyd et al. (Lloyd et al., 2016).

Multimodal physiological data recordings from 25 preterm infants with a median GA of 28 weeks (IQR: 26 – 29 weeks) at the NICU of Cork University Maternity Hospital, Ireland were used in this study. Clinical characteristics of preterms in the dataset are provided in Table 5.1. The dataset includes continuous, synchronously recorded, ECG, EEG and BP signals. The duration of recordings used in this study totals 872 hours (median = 37 hours, IQR: 24 to 48 hours). The duration and temporal location of recordings are represented in Figure 5.2. Multi-channel EEG was recorded using a Natus NicOne video EEG machine using the international 10-20 system of electrode placement, adjusted for neonates, with the analysis performed on the 8 bipolar channels: F4–C4, C4–O2, F3–C3, C3–O1, T4–C4, C4–Cz, Cz–C3, and C3–T3.

Table 5.1: Clinical information for preterms in the dataset.

Subject #	GA (weeks)	BW (g)	Gender	Apgar score 5 min	Umbilical cord pH	CCS
1	30	1540	F	8	7.1	0
2	28	980	F	10	7.26	1
3	30	1450	M	5	7.02	1
4	29	1230	M	9	7.27	0
5	26	840	M	8	7.23	0
6	25	640	F	6	7.34	1
7	31	960	F	9	7.23	0
8	29	1460	F	-	6.67	1
9	30	1000	F	10	7.24	0
10	28	980	F	8	7.16	0
11	28	650	F	7	7.22	0
12	28	530	F	6	7.16	0
13	26	860	M	8	7.12	1
14	26	980	M	8	7.24	1
15	24	740	F	9	7.18	1
16	24	670	M	6	6.84	1
17	28	1040	F	9	7.3	0
18	23	540	F	7	7.22	1
19	26	660	F	3	7.11	1
20	28	1330	F	4	7.08	0
21	30	730	M	10	7.32	0
22	23	580	F	6	7.24	1
23	28	680	M	8	7.32	1
24	28	1130	F	9	7.15	0
25	31	1900	M	8	7.02	0
Median (IQR)	**28 (26 to 29)**	**960 (670 to 1130)**	**64% (F)**	**8 (6 to 9)**	**7.22 (7.11 to 7.24)**	**48 % (1 - sick)**

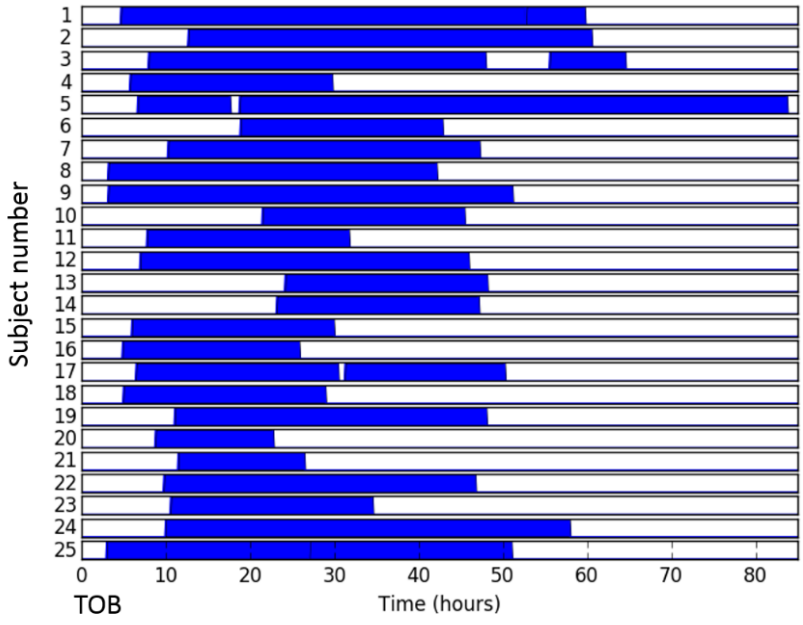


Figure 5.2: Schematic representation of duration and temporal location of recordings. Each recording is represented with respect to the time of birth (TOB) for each neonate.

This system also recorded continuous single channel ECG. EEG and ECG were sampled at 256 Hz. Continuous invasive arterial BP monitoring was simultaneously performed via an umbilical arterial catheter using the Philips Intellivue MP70 machine, which provides BP data sampled at 1 Hz. The synchronously recorded data of each patient was exported and stored locally as a single file. An example of a data segment is presented in Figure 5.3.

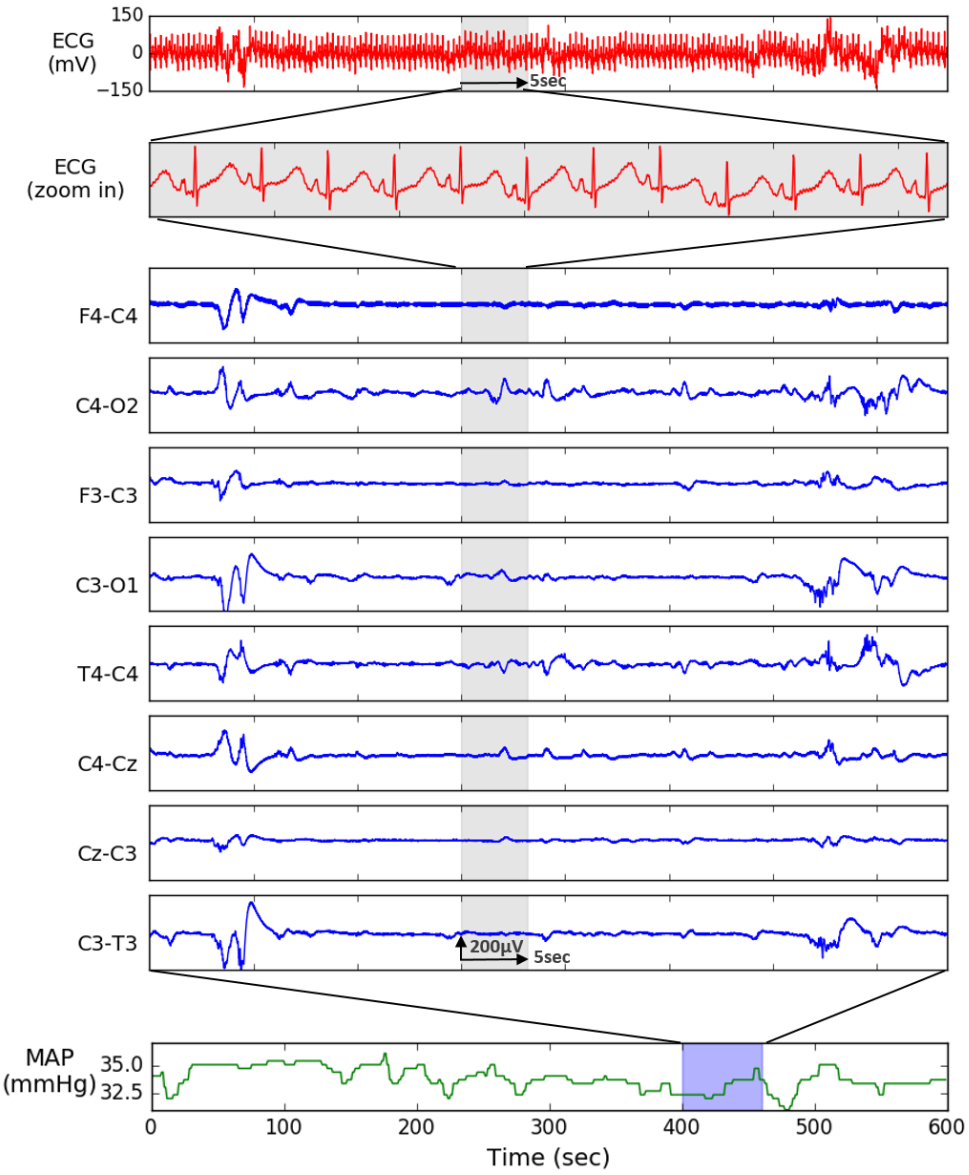


Figure 5.3: One minute of raw ECG and eight-channel EEG and ten minutes of mean arterial pressure (MAP) recordings (GA=26 weeks).

A recent survey has confirmed that the GA-based rule remains the most common criterion used by neonatologists to direct intervention for hypotension management (Stranak et al., 2014). In this study, we aim to investigate the effect of hypotension on preterm wellbeing. The term 'hypotension' in this work does not confer a clinical diagnosis, but here simply refers to episodes when the BP (MAP) falls below some specified age-related (GA) threshold: MAP ≤

GA, MAP ≤ GA + 2 mmHg and MAP ≤ GA + 4 mmHg. Every threshold was represented by a subset of preterm neonates based on the criteria that during the recording there was at least one episode of at least 5 min duration where MAP fell below a given threshold. For the selected MAP thresholds: GA, GA+2 mmHg, and GA+4 mmHg, the resultant EEG datasets contained 18 (9 healthy), 22 (11 healthy) and 25 (13 healthy) subjects, respectively. The ECG datasets was represented with 15 subjects (8 healthy) for the threshold equal to GA, 19 subjects (10 healthy) for the threshold equal to GA + 2 mmHg, and 23 subjects (12 healthy) for the threshold equal to GA + 4 mmHg. In both cases, for EEG and ECG data, the number of subjects obviously reduces with the tighter thresholds. Segments which were highly corrupted by artefacts were excluded from the study which is why the number of subjects in the ECG and EEG datasets for the same threshold is different. In addition, for each qualifying recording, the episodes under the chosen threshold were marked as episodes of hypotension rather than marking the whole recording as hypotensive. This was done in order to assess if HRV and EEG characteristics are affected by hypotensive episodes.

5.3 Feature extraction

The ECG signal was segmented into non-overlapping 5-minute epochs (Goulding et al., 2015). The R peaks were identified using the Pan-Tompkins method (Pan and Tompkins, 1985) as explained in Chapter 3. The RR intervals were used to estimate the instantaneous HR signal. Abnormal values of the time intervals between R peaks (RR intervals) caused by artefacts were corrected using a moving average filter or discarded if the epoch was too corrupted. The corrected RR intervals, NN intervals, were used to estimate the instantaneous heart rate signal. The behaviour of HR was quantified using time-domain, frequency-domain, and nonlinear HRV analysis. A concise list of the thirteen HRV features is presented in Table 5.2.

Prior to EEG feature extraction, the amplitude-based thresholding of the EEG was performed for every channel to remove zero-signal and high amplitude artefact (e.g. eye blinking, DC, moved/disconnected electrode). The EEG signal was filtered to the range of 0.3-30 Hz and down-sampled to 64 Hz. The signal in each channel was segmented into 1-minute epochs with a 1-second shift. Time domain EEG features used in this study included activity, mobility, complexity (Hjorth, 1970) and the number of zero crossings. Frequency domain features were calculated in different EEG frequency bands: 0.3–3 Hz, 3–8 Hz, 8–15 Hz and 15–30 Hz. For each sub-band, the absolute and relative power (normalized by the total power) were calculated. The spectral entropy (SE) feature was also extracted for each sub-band. For every EEG feature, the median value across all eight channels was calculated. A detailed explanation of the feature extraction procedure is provided in Chapter 3. Sixteen EEG features are listed in Table 5.2.

Table 5.2: Frequency- and time-domain features extracted from EEG, ECG, and BP.

Domain	HR features
Time	MeanRR, SDNN, skewness, kurtosis, TINN, RMSSD, SDNN/RMSSD, ApEn, Allan Factor
Frequency	Power in VLF (0.008 – 0.04 Hz), LF (0.04 – 0.2 Hz) and HF (0.2 - 1 Hz) bands, ratio LF/HF
	EEG features
Time	Activity, mobility, complexity, zero crossing
Frequency	Relative power (RP), total power (P) and spectral entropy (SE) in 0.3–3 Hz, 3–8 Hz, 8–15 Hz and 15–30 Hz bands
	BP features
Time	MAP, MAP-GA, MAP/GA

5.4 Exploring the predictive power of HRV characteristics

The area under the receiver operating characteristic curve (AUC) is used to quantify the discriminative (predictive) power of each individual HRV and EEG feature with respect to the health status of the preterm represented by CCS. This allowed us to investigate how features are related to the problem of interest.

The class-specific histograms (healthy and unhealthy neonates) of the HRV characteristics used in this study are represented in Figure 5.4. It can be seen that no clear separation between the two classes can be observed. At the same time histograms of some characteristics, such as skewness and Allan factor, are highly overlapped.

In order to investigate the effect of low BP on the preterm wellbeing, the HRV characteristics were studied for different levels of MAP. The predictive power of each HRV feature is presented in Table 5.3. The AUC of each HRV feature computed on **All epochs** extracted from the whole recording was contrasted with the AUC which was computed for epochs during the episodes of hypotension (**Hypotensive events**). This has been carried out for 3 different thresholds (resulting in Set 1, 2 and 3) as explained in Section 5.2.

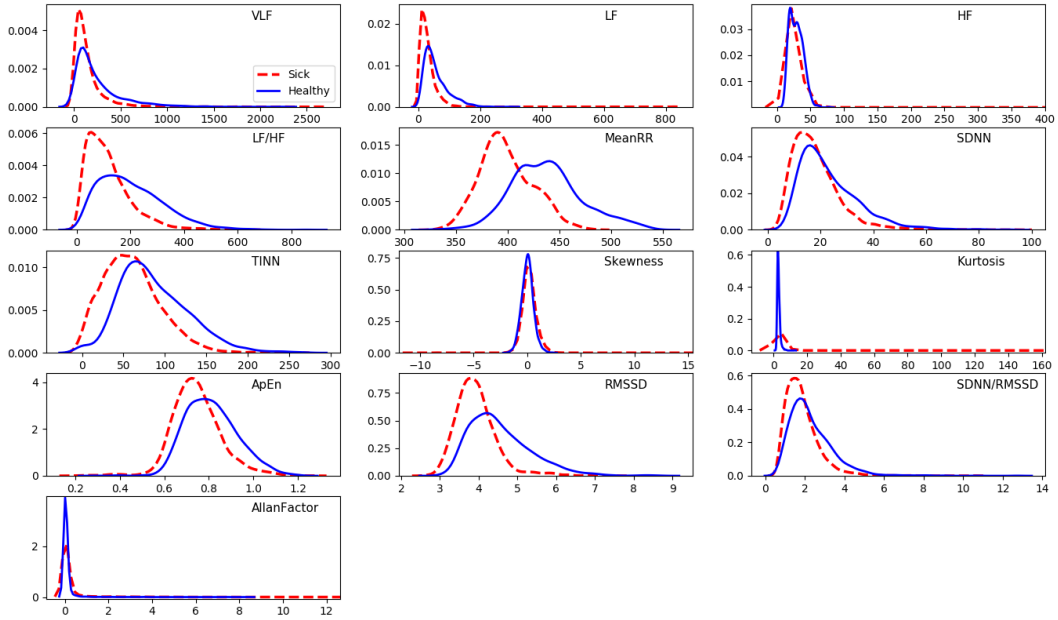


Figure 5.4: Class specific histograms of the thirteen features extracted from all HRV epochs of 23 preterm neonates. No distinctive separation can be observed between the two classes.

Table 5.3: The predictive power of the HRV features measured using AUC. 'All epochs' represents complete recordings. 'Hypotensive events' represents only epochs under the specific MAP threshold (GA, GA+2 or GA+4) for the same set of babies. Δ represents a change (an increase or decrease) of the AUC.

HRV features	Set 1 MAP ≤ GA, 15 subj			Set 2 MAP ≤ GA+2, 19 subj			Set 3 MAP ≤ GA+4, 23 subj		
	All epochs (3968)	Hypotensive events (316)	Δ	All epochs (5282)	Hypotensive events (779)	Δ	All epochs (6217)	Hypotensive events (1488)	Δ
VLF	0.66	0.86	↑	0.70	0.77	↑	0.67	0.70	↑
LF	0.74	0.89	↑	0.78	0.81	↑	0.76	0.78	↑
HF	0.65	0.90	↑	0.68	0.84	↑	0.64	0.76	↑
LF/HF	0.69	0.73	↑	0.72	0.68	↓	0.73	0.69	↓
MeanRR	0.82	0.84	↑	0.78	0.83	↑	0.82	0.85	↑
SDNN	0.65	0.85	↑	0.68	0.80	↑	0.65	0.73	↑
TINN	0.7	0.89	↑	0.73	0.83	↑	0.70	0.77	↑
Skewness	0.52	0.60	↑	0.56	0.51	↓	0.58	0.52	↓
Kurtosis	0.57	0.58	↑	0.57	0.61	↑	0.56	0.61	↑
ApEn	0.70	0.92	↑	0.69	0.85	↑	0.67	0.78	↑
RMSSD	0.77	0.97	↑	0.77	0.93	↑	0.76	0.87	↑
SDNN/ RMSSD	0.61	0.71	↑	0.66	0.68	↑	0.64	0.65	↑
AllanFactor	0.52	0.66	↑	0.52	0.62	↑	0.51	0.55	↑

The obtained results indicate that HRV is sensitive to BP and the features extracted during the episodes of low MAP improve the prediction of the short-term outcome for the preterm neonates. From Table 5.3 it is clear that many HRV features, such as RMSSD, ApEn and HF, are relevant for the task and bear moderate predictive power with respect to the outcome. Examining HRV during episodes of low BP improves the predictive power of every relevant

HRV feature in all three Sets. For example, in Set 3, RMSSD improved from AUC of 0.76 to 0.87, in Set 2 - from AUC of 0.77 to 0.93, and in Set 1 - from AUC of 0.77 to 0.97. The improvement increases with tighter definitions of **Hypotensive events** through the GA-based threshold. This indicates the sensitivity of HRV to changes in BP. The tighter the threshold, the larger the effect of lower BP and the larger the improvement in AUC that is observed. However, this comes at the cost of a reduced number of subjects and thus reduced confidence in the observed changes for tighter thresholds. It is worth noting that each Set was restricted to include only those newborns with the presence of hypotensive events as defined by the corresponding GA-based thresholds. For each Set, the segments with hypotensive events come from the same subjects and represent a subset of the category **All epochs**. Thus, the obtained improvement is purely attributable to the effect that BP has on HRV rather than an effect which could have been caused by including extra subjects in the category of **All epochs**.

From Table 5.1 it can be seen that 10 out of the 12 abnormal outcomes are present in infants ≤ 28 weeks GA. Figure 5.5 shows the probability density functions (PDF) of the RMSSD feature for the **All epochs** dataset, the subset with normal BP and the subset of **Hypotensive events** for the cohort of preterms with both healthy and unhealthy outcomes. These PDFs are demonstrated for 1) the complete dataset (All GA, 23 subjects), 2) GA>28 subset (6 subjects, 1 poor outcome), and 3) GA≤28 subset (17 subjects, 10 poor outcomes). All distributions for hypotensive events demonstrate a shift in the distributions towards the right for the healthy neonates. This indicates that unhealthy infants of all GAs, (not just extremely preterm infants with GA≤ 28 weeks), cannot alter their HRV in response to hypotension, as the distributions remain largely undisturbed.

The separation between distributions increases when considering only hypotensive episodes, Figure 5.5 (All GA, c). Comparing the Figure 5.5 (b) and Figure 5.5 (c) it can be seen that the HRV of healthy newborns reacts to episodes of low BP by increasing its median value from *4.21 (IQR: 3.9 to 4.6)* under normal BP to *4.96 (IQR: 4.2 to 5.5)* under **Hypotensive events**. At the same time, for unhealthy newborns, no such increase is observed, with the median RMSSD value of *3.89 (IQR: 3.6 to 4.2)* under normal BP and *3.75 (IQR: 3.5 to 4.1)* under **Hypotensive events**. This indicates that the HRV of healthy preterms reacts to the drop in BP. In (Semenova et al., 2018) the increased level of interaction between EEG and BP in preterms was shown to be associated with lower risks of illness severity. Similarly here, a strong interaction between HRV and BP is associated with a good outcome, whereas the lack of interaction is associated with a poor outcome.

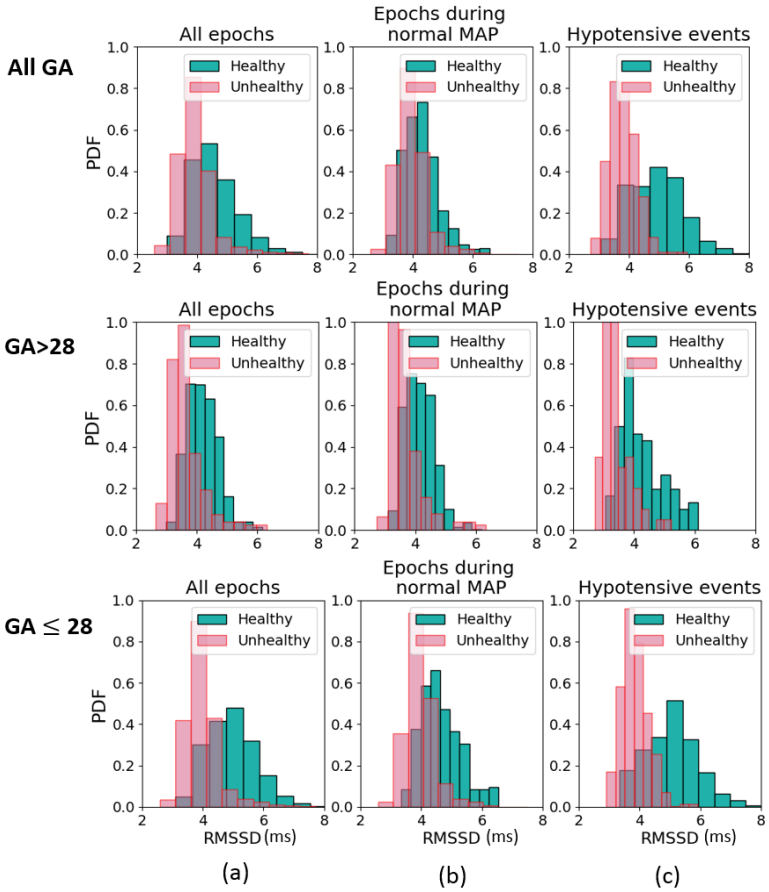


Figure 5.5: PDF for the RMSSD feature. Original subset (a) contains RMSSD feature values from the complete recordings. Normal (b) and hypotensive (c) subsets represent RMSSD feature extracted during episodes of normal BP (MAP > GA+4) and during hypotensive events (MAP ≤ GA+4). PDFs are demonstrated for all 23 subjects (All GA), GA>28 subset (6 subjects) and GA≤28 subset (17 subjects).

Figure 5.6 shows a comparison between healthy and unhealthy neonates for several relevant HRV features. The obtained results indicate that preterms with abnormal outcome have significantly lower values of HRV ($p<0.001$) even for **All epochs** dataset. This separation increases for **Hypotensive events** as indicated in Table 5.3 by the increased AUCs.

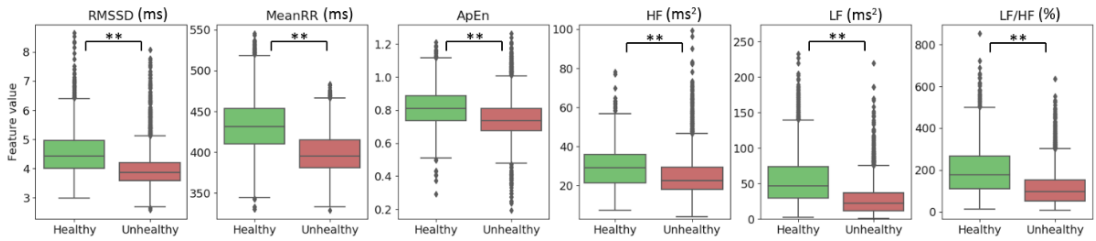


Figure 5.6: Values of HRV features extracted from All epochs dataset for healthy and unhealthy preterm neonates. Boxplot analyses show the median, 25th and 75th percentiles, and the outliers. '**' represent statistically significant differences between groups with p<0.001 using Mann-Whitney U test. The predictive power of the features quantified by AUC is presented in Table 5.3 (All epochs, Set 3).

The 3D projection of the first 3 principal components obtained using principal component analysis (PCA) applied on the **All epochs** dataset and **Hypotensive events** subset is demonstrated in Figure 5.7. The combination of multiple features improves the separation in both cases, with very little overlap observable for **Hypotensive events**.

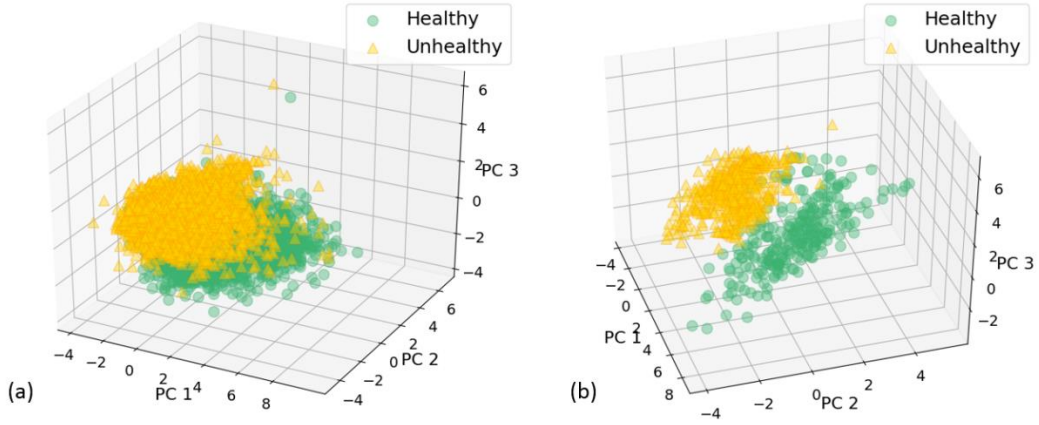


Figure 5.7: Principal component analysis (PCA) of the dataset comprised of all epoch (a) and epoch during hypotensive episodes (MAP ≤ GA+4) (b). In this study, PCA is used as a tool for exploratory feature analysis which is aimed at checking the discriminative power of the HRV feature set with respect to the short-term outcome of the preterm neonate.

The most important HRV feature, RMSSD, estimates the short-term component of HRV and represents the parasympathetic activity of the heart. In (Dimitrijević et al., 2016) RMSSD was shown to improve 2-year outcome prediction for preterm neonates, with lower RMSSD values corresponding to minor neurologic dysfunction and cerebral palsy. The work presented in this thesis has also shown that RMSSD values are lower (Figure 5.5) for unhealthy preterm neonates. Other studies have reported that the RMSSD feature is a good early predictor of a septic shock for adults (Chen and Kuo, 2007) and sudden unexplained death in epilepsy (DeGiorgio et al., 2010). Decreased HRV was previously associated with hypoxic brain injury for newborns (Matić et al., 2013) and with the failure of the first extubation for the preterm infants (Kaczmarek et al., 2013). Other HRV characteristics which are shown to be relevant to the short-term outcome of the preterm neonate (Figure 5.6) were previously reported to be indicative of neonatal health status. The ApEn measure was proposed by Pincus (Pincus, 1991) who reported reduced ApEn in distressed foetuses (Pincus and Viscarello, 1992) and sick newborns (Pincus et al., 1991). This can be interpreted as an increased regularity of cardiac rhythm. Further application of entropy estimators to neonatal HRV have revealed an association with neonatal sepsis (Lake et al., 2002). In this study, we have obtained similar results (Figure 5.6) with lower ApEn values obtained for unhealthy preterms. The TINN characteristic was previously shown (Goulding et al., 2015) to be significantly associated with the severity of the HIE for full-term neonates, with lower TINN values corresponding to sick

neonates. Similarly, lower TINN values in this study are shown to be associated with an increased risk of morbidity of the preterm.

Neonatal HRV evaluated by spectral analysis is usually characterised by the dominant activity in the LF band. LF is mediated predominantly by the sympathetic component of the autonomic nervous system. LF of HRV is also considered to represent Mayer waves of BP changes (Draghici and Taylor, 2016). Calculation of the LF/HF ratio is a method to establish the ratio between the components of the autonomic nervous system – sympathetic/parasympathetic balance (Electrophysiology, 1996). Reduced LF and HF features were reported for neonates with HIE, implying the reduction of autonomic function (Goulding et al., 2015). Similar results were observed in the current study with a decreased power in the LF and HF bands (Figure 5.6). In general, both human and animal studies have supported the finding that reduced spectral power in the LF component is indicative of the impaired function of the autonomic nervous system (Piccirillo et al., 2009; A. J. Shah et al., 2013).

An example of the association between the MAP and two best HRV features, MeanRR and RMSSD, for one healthy and one unhealthy neonate is represented in Figure 5.8 using linear regression. It can be seen that observed unhealthy preterm is not able to compensate for the changes in MAP. This is represented by a narrow range of the values of HRV characteristics for both high and low levels of the MAP. It indicates that the HRV of the sick neonate has very minor or no reaction to the fluctuations (decrease and increase) in the MAP. At the same time, a healthy neonate reacts to the changes in MAP. This is represented by the wider range of HRV characteristics for different levels of the MAP. The correlation between MeanRR and MAP for the healthy and sick neonates is $r = 0.32$ and $r = 0.15$ correspondingly; and between RMSSD and MAP: $r = 0.2$ and $r = 0.14$. The test on the significance of the difference between two correlations (one for unhealthy and one for healthy group of neonates) has shown that correlations were significantly different for the MeanRR and MAP variables ($p = 0.017$) (Figure 5.8 (a)). At the same time, no significant difference was found between correlations of healthy and unhealthy neonates for RMSSD and MAP (Figure 5.8 (b)).

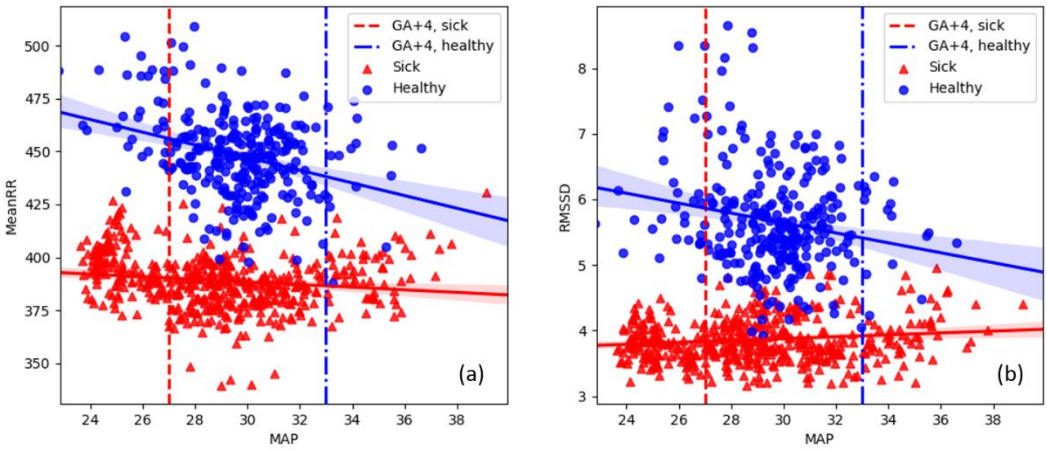


Figure 5.8: An example of the association of the MAP with MeanRR and RMSSD features for one healthy (GA=29 weeks) and one unhealthy (GA=23 weeks) neonate. Shaded regions correspond to 95% CI. The correlation between MeanRR and MAP for the healthy and sick neonates is r = 0.32 and r = 0.15 correspondingly; and between RMSSD and MAP: r = 0.2 and r = 0.14.

5.5 Exploring the predictive power of EEG characteristics

The EEG characteristics used in this study were chosen according to the prior knowledge about the maturational changes taking place within the preterm brain, as well as the findings of the previous studies conducted on the population of preterm neonates. Figure 5.9 represents histograms of all EEG features for both healthy and unhealthy neonates.

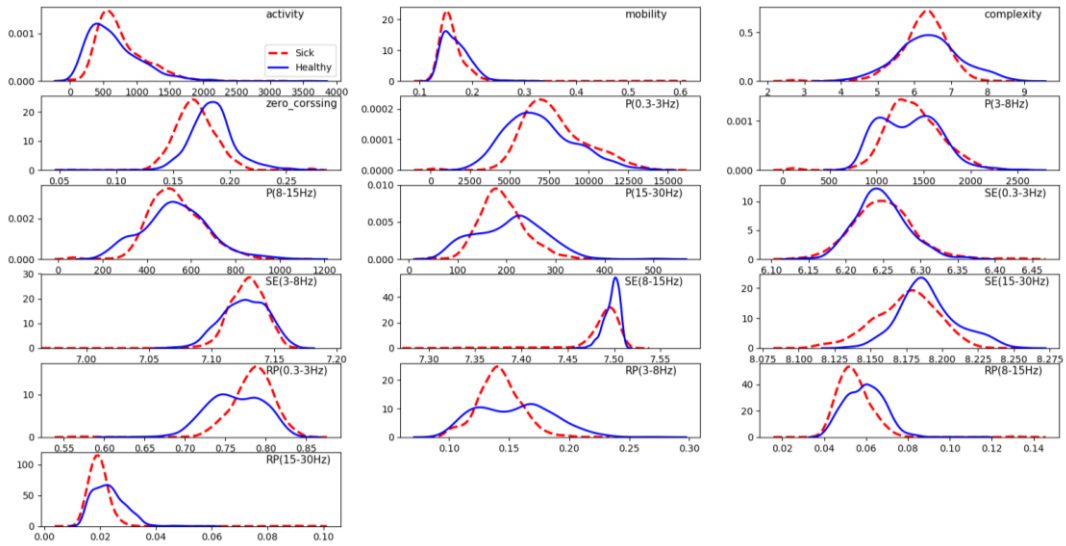


Figure 5.9: Class specific histograms of the EEG features extracted from all epochs of 25 preterm neonates; no clear separation can be observed between healthy and unhealthy neonates.

The predictive power of various EEG features using previously defined MAP thresholds (GA, GA+2, and GA+4) is summarized in Table 5.4.

Table 5.4: The predictive power of the EEG features measured using AUC. 'All epochs' represent complete recordings. 'Hypotensive event' represents only epoch under the specific MAP threshold (GA, GA+2 or GA+4) for the same set of babies. Δ represents a change (an increase or decrease) of the AUC.

EEG feature	Set 1 MAP ≤ GA, 18 subj			Set 2 MAP ≤ GA+2, 22 subj			Set 3 MAP ≤ GA+4, 25 subj		
	All epochs (7688)	Hypotensive events (800)	Δ	All epochs (9580)	Hypotensive events (1510)	Δ	All epochs (10444)	Hypotensive events (2564)	Δ
P (0.3-3 Hz)	0.55	0.68	↑	0.58	0.67	↑	0.58	0.65	↑
P (3-8 Hz)	0.57	0.55	↓	0.52	0.54	↑	0.53	0.52	↓
P (8–15 Hz)	0.60	0.65	↑	0.54	0.60	↑	0.55	0.53	↓
P (15–30 Hz)	0.58	0.67	↑	0.59	0.62	↑	0.61	0.55	↓
RP (0.3-3 Hz)	0.69	0.68	↓	0.64	0.71	↑	0.64	0.70	↑
RP (3-8 Hz)	0.64	0.77	↑	0.57	0.73	↑	0.57	0.69	↑
RP (8-15 Hz)	0.66	0.57	↓	0.63	0.51	↓	0.66	0.60	↓
RP (15-30 Hz)	0.62	0.69	↑	0.67	0.58	↓	0.69	0.53	↓
SE (0.3-3 Hz)	0.56	0.78	↑	0.57	0.68	↑	0.56	0.64	↑
SE (3-8 Hz)	0.53	0.65	↑	0.57	0.64	↑	0.51	0.60	↑
SE (8–15 Hz)	0.68	0.78	↑	0.73	0.76	↑	0.70	0.75	↑
SE (15–30 Hz)	0.63	0.54	↓	0.6	0.59	↓	0.60	0.62	↑
Activity	0.51	0.66	↑	0.54	0.65	↑	0.54	0.62	↑
Mobility	0.64	0.53	↓	0.6	0.54	↓	0.62	0.58	↓
Complexity	0.50	0.67	↑	0.57	0.60	↑	0.56	0.57	↑
Zero crossing	0.61	0.51	↓	0.66	0.53	↓	0.66	0.56	↓

Obtained results indicate that not all EEG features are relevant for the task of short-term outcome prediction. Spectral entropy (8-15 Hz), relative power in 0.3-3 Hz and 3-8 Hz sub-bands are the top three EEG features with the highest discrimination power. Similarly to HRV characteristics, no clear separation between healthy and unhealthy neonates can be observed when considering EEG features for all range of MAP (Figure 5.9). However, it can be also appreciated that the AUCs for these features have similarly increased for **Hypotensive events**. When tightening MAP threshold and considering EEG characteristics for low levels of BP, the corresponding AUC values for the top three EEG features have improved from 0.7 to 0.75, from 0.64 to 0.7 and from 0.57 to 0.69 respectively (Table 5.4, Set 3).

The results of the study indicate that EEG features have lower discriminative power in regards to the short-term outcome with a maximum AUC of 0.75 achieved by spectral entropy (8-15 Hz) feature as compared to an AUC of 0.87 for the RMSSD feature extracted from ECG. A comparison between the predictive powers of the EEG and HRV features is shown in Figure 5.10. The figure clearly demonstrates that the HRV characteristics are more sensitive to changes in MAP as the performance increases more abruptly with the tightening of the threshold on the definition of hypotensive episodes.

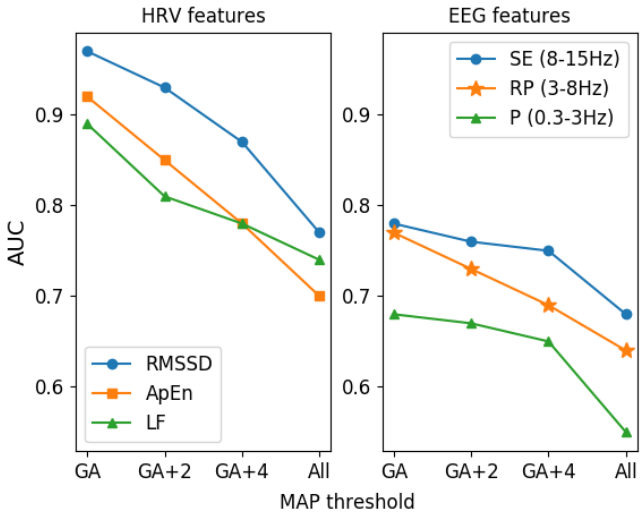


Figure 5.10: AUC of three EEG and HRV features for various thresholds and the **All epochs** dataset.

The lower predictive power of the EEG could be caused by a number of reasons. EEG is known to be a complex nonlinear signal, which might be difficult to relate to the outcome of the preterm (Semenova et al., 2017) using linear statistical assessment based on the AUC; this is clearly different to HRV, which has previously been successfully associated with preterm health outcomes (Cabal et al., 1980; Doheny et al., 2014). Likewise, it is possible that BP may not be directly related to cerebral function. More specifically, it should be noted that EEG activity reflects end organ activity and therefore may be influenced by other unaccounted for parameters. In comparison, it is not a surprise that HRV should be more closely linked to BP via the autonomic system. Cerebral oxygenation measured using NIRS is another surrogate measure of brain activity, which is more closely related to cerebral blood flow and BP (Lou et al., 1979; Ment et al., 1984).

In (Ahmed et al., 2016) 1-hour-long EEG recordings were used for the grading of HIE severity in term neonates; this might suggest that lower predictive power of EEG could have been caused by the choice of the decision-making window of 5 min duration, which is too short to capture the information related to outcome prediction. Similarly, hypotensive episodes were generally too short to allow the capture of information related to the brain health of the preterm from EEG.

As shown in Figure 5.1 physiological signals were recorded during first 72 hours of life only. After this point in time until the discharge from the NICU, preterm neonate is exposed to various risks such as IVH, necrotizing enterocolitis, retinopathy of prematurity and other morbidities. This implies that the CCS at discharge could also be influenced by a number of possible complications. More specifically, if the preterm has suffered from IVH during the

stay in the NICU at any point after the first 72 hours of life, the EEG recordings used in this study would not capture the possible abnormal activity related to the haemorrhage.

Previously EEG was successfully used for the long-term neurodevelopmental outcome prediction for both term and preterm infants (Lloyd et al., 2016; Sinclair et al., 1999). The EEG characterizes neurological function, whereas the clinical score used in the study represents a general health status of the preterm. HRV, on the other hand, characterizes the current physiological stability and therefore may be considered as a more suitable method for the prediction of short-term health status.

5.6 Machine learning: boosted decision trees

There exist a number of different supervised ML techniques and many of them have been successfully applied in the medical field. The choice of the ML method depends on a number of factors, namely, a problem that needs to be solved, the type of data available, as well as the interpretability of the obtained results. Unlike NN for example, which require normalisation /rescaling of input features, the tree-based models are easy to use as they are invariant to the input scale and also accommodate categorical and missing data.

When it comes to clinical decision making, the derivation of the reasoning from a constructed model is of great importance. This information can provide additional insight into the understanding of the physiology, clinical condition as well as the possible connection of the processes taking place within a human body. Therefore, in this study for the problem of the neonatal health outcome prediction based on the combination of the EEG and HRV features a tree-based model classifier was used. The overall diagram of the multimodal short-term outcome prediction is represented in Figure 5.11.

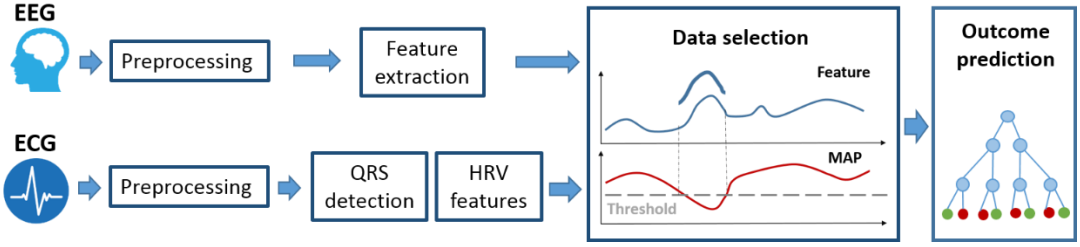


Figure 5.11: The overall diagram of the multimodal short-term outcome prediction using a boosted decision tree classifier. Mean arterial pressure (MAP) threshold, is used as a data selection technique.

5.6.1 Tree models: ensemble methods

A brief introduction to tree models is provided in Chapter 3. Decision trees learn the decision boundary by recursively partitioning the feature space into non-overlapping regions using

some defined threshold. An example of a nonlinear decision boundary constructed with a decision tree using two HRV feature is represented in Figure 5.12.

In most problems, adding a large number of trees, beyond a certain limit, does not improve the performance of the classifier. This is due to the fact that each new tree is constructed in order to correct for the errors made by the sequence of previous trees. As a result, at some point, the classifier stops reducing the classification error. This is visually demonstrated in Figure 5.12, using decision boundaries which were constructed using 50, 300 and 1000 decision trees. Obtained results show no apparent difference between the decision boundaries constructed with 300 and 1000 trees. This suggests that for a particular example, 300 trees may be a sufficient number to generate a decision. The decision regarding the number of trees to be used is usually optimised during the cross-validation stage, along with other parameters.

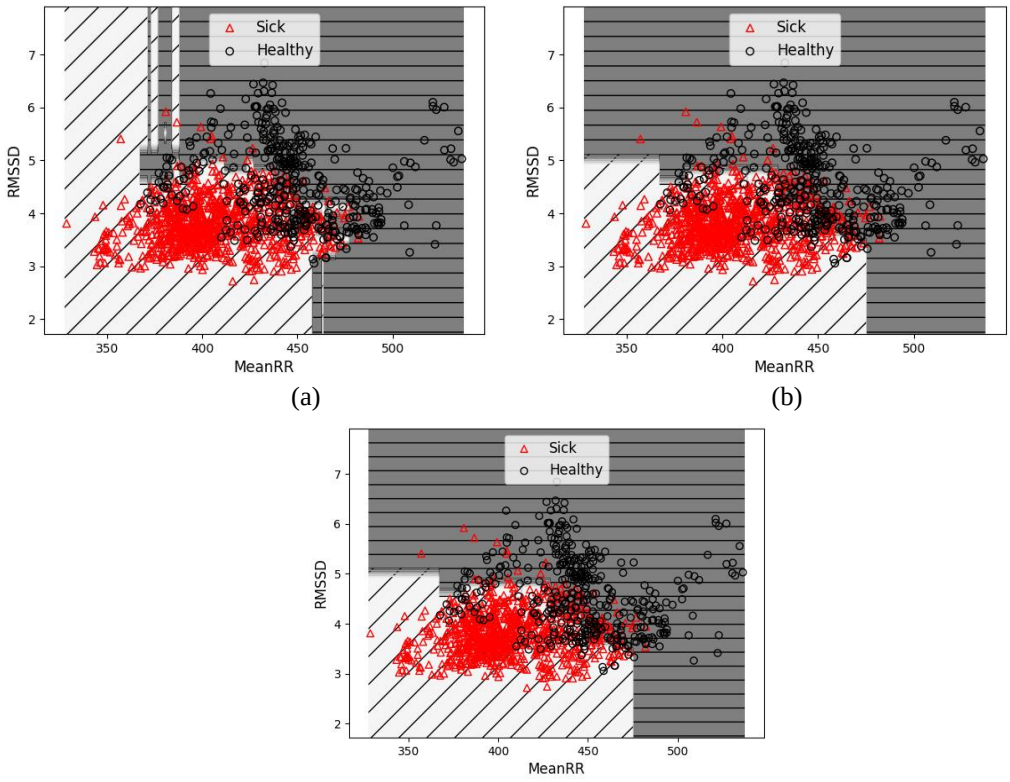


Figure 5.12: An example of the nonlinear decision boundary constructed with boosted decision tree classifier using two HRV features with (a) 50 trees, (b) 300 trees and (c) 1000 decision tees. It can be seen that the decision boundary constructed with 50 trees (a) differs from the one constructed with 100 trees. At the same time, no apparent distinction can be observed between 300 (b) and 1000 (c) trees.

Usually, in order to make a good prediction, one tree is not enough. As a result, for an accurate prediction with tree-based models, a tree ensemble is used. Ensemble methods allow using multiple learning algorithms to get better predictive performance (Rokach, 2010). A general diagram that represents the ensemble classification framework is provided in Figure 5.13.

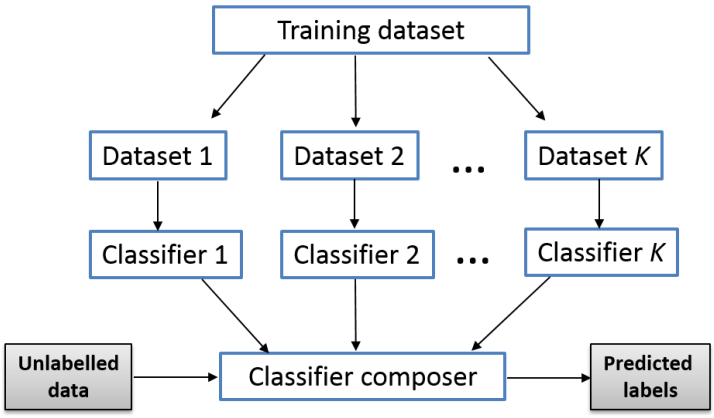


Figure 5.13: A general diagram of the ensemble classification framework.

There are two main types of the ensemble framework, namely, dependent and independent ways of building ensembles. The framework, where the output of the classifier is used for the construction of the next one is referred to as a dependent framework. This type of framework allows guiding further decision based on the knowledge generated by the previous interaction. As a result, a new classifier is generated to comprehend the examples for which the previous tree did not do well. In the independent framework, each classifier is built independently and their corresponding outputs are then combined to generate a single output.

Independent methods

Using the independent methodology of the ensemble classification, the original dataset is divided into several smaller datasets (sub-sets), based on which classifiers are trained. The data in sub-sets may be overlapped or mutually exclusive. This is followed by a classifier composer step after which a final decision is generated.

Bagging, also known as a bootstrap aggregating method, is a well-known example of the independent ensemble method (Breiman, 1996), where each classifier is trained on a training set uniformly sampled with replacement from the original dataset. Sampling with replacement implies that some observations may be repeated more than once, while others may not be included at all. The composite classifier generates a final output by averaging the output for regression problems, or majority voting for classification tasks. If perturbation of the training set causes significant changes in the model, such a classifier is then referred to as unstable. Bagging procedures have been demonstrated to improve unstable classifiers such as artificial neural networks, classification and regression trees (Breiman, 1996).

A random forest is an example of an independent ensemble classification (Breiman, 2001). Random forest generates predictions based on the combination of a large number of unpruned decision trees. There exist different ways to construct random forest tree ensembles. It was shown that some random forests tend to have consistently lower generalization error. This was

achieved by the following techniques: introducing random noise into the outputs (Breiman, 2000); random split selection, where each split is randomly selected from a number of best splits (Dietterich, 2000); and using a random selection of features to grow each tree (Ho, 1998).

Dependent methods

Boosting is a method for the creation of an accurate and strong classifier from a set of weak classifiers (Schapire, 2003). AdaBoost is a well-known example of the dependent ensemble classification (Freund and Schapire, 1996). The main idea behind this algorithm lies in giving more focus, quantified by a weight assigned to instances which are harder to classify. After each iteration, the misclassified instances gain more weight, while the weight of the correctly classified example is decreased. Each individual tree (classifier) is then weighted according to its accuracy. As a result, trees which perform more accurate classification have higher weights and consequently have a higher contribution to the final prediction. The final prediction is obtained by voting on the weighted classifiers.

Gradient boosting is another powerful algorithm. While AdaBoost uses up-weighting of the misclassified instances, gradient boosting identifies misclassification from the large residual obtained on the previous iteration. It regulates the importance of the misclassified instances by training weak classifiers on the remained errors (residuals) of a strong classifier. The residuals are computed after each iteration. The process is carried out until the residuals are close to zero. The contribution of each weak classifier to the strong one is computed using the gradient descent optimization technique. This allows for the computation of the contribution which minimizes the error of the strong classifier.

The gradient boosting procedure is shown in Figure 5.14, where the decision tree model is fitted to the simulated data with input X and output Y. On each stage, the residuals are calculated and a new model is then represented as s sum of the previous model and the model which is obtained by fitting to residuals. Figure 5.14 demonstrates predictions and the residuals of the classifier which are computed for a number of iterations. It can be seen that after each iteration, the residuals decrease and eventually shrink to the values close to zero. At the same time, the predictions gradually improve and after fifteen iterations the predicted values are very close to the ground truth. At this point the training can be stopped; this will allow avoiding overfitting by building a very complex model. For the example in Figure 5.14, the model clearly overfits after 50 iteration, where the classifier tries to fit each training example.

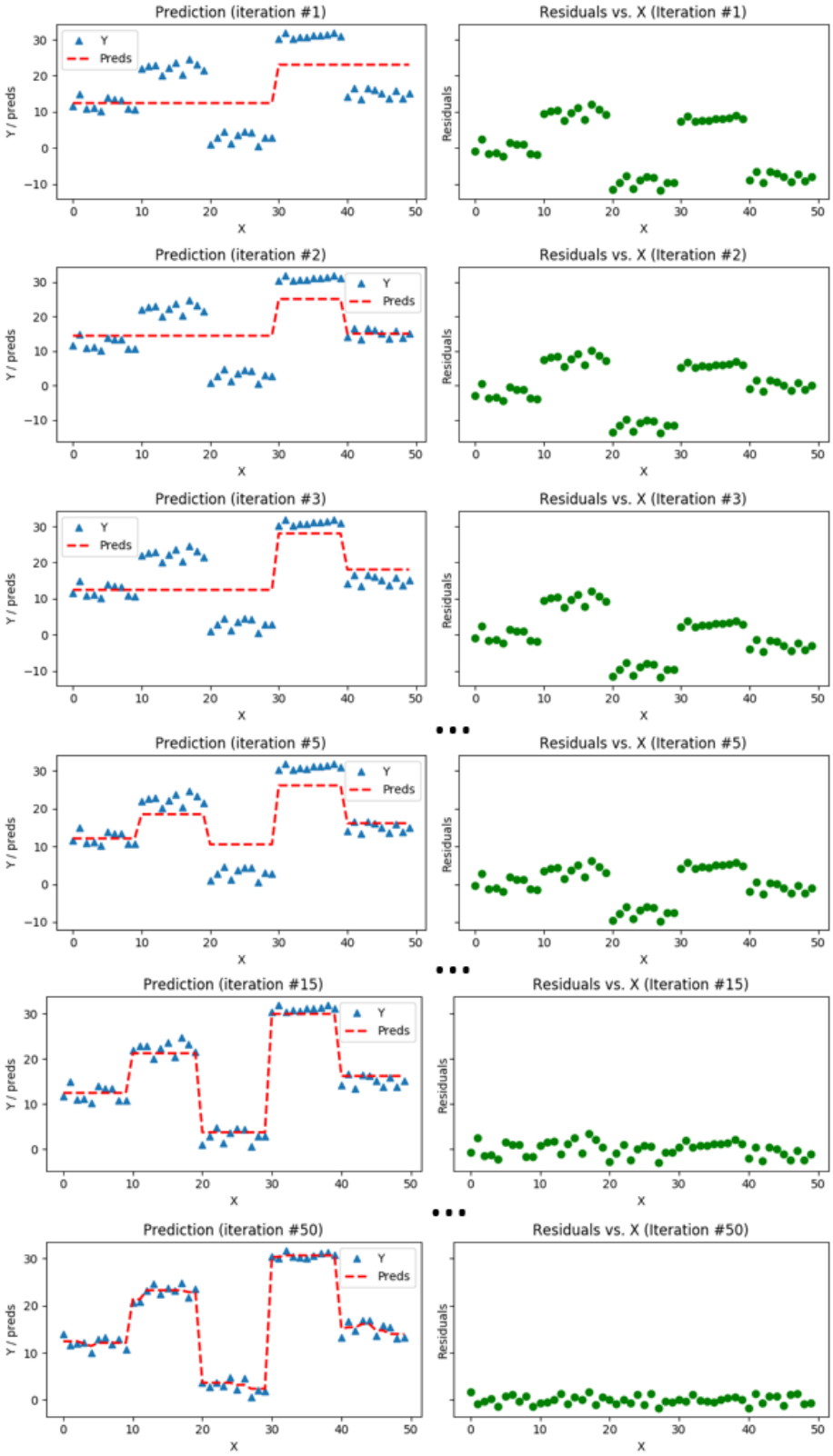


Figure 5.14: Visualisation of gradient boosting predictions using simulated data.

5.6.2 Gradient boosting with regularization

Dependent ensemble methods are known to outperform the independent ones (Breiman, 2001). An extended and even more powerful version of gradient boosting is boosting with

regularisation. The extreme gradient boosting classifier, XGBoost (Chen and Guestrin, 2016) is a well-known implementation of boosted decision trees with regularization which is available as open source software. The package has been widely recognized in a number of ML and data mining challenges, (e.g. in Kaggle competitions), where state-of-the-art results on a wide range of problems have been reported.

Given a set of *N* training examples of the form $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ such that $x_i \in R^d$ is a *d*-dimensional feature vector of the *i*-th example and y_i is its label; the goal is to create a model, f, to predict values in the form:

$$\hat{y} = \sum_{k=1}^{K} f_k(x) \tag{5.1}$$

At each stage of gradient boosting, $1 \leq k \leq K$, a model, f_k, is constructed. A new improved model is then constructed, $f_{k+1}(x)$, that adds an estimator, *h*, as:

$$f_{k+1}(x) = f_k(x) + h(x) \tag{5.2}$$

by fitting *h* to the residuals, $(y - f_k(x))$. More formally, the following objective function is minimized:

$$Obj = \sum_{i=1}^{N} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{5.3}$$

Here l is a differentiable convex loss function that measures the difference between the prediction $\hat{y}_i$ and the target y_i; each f_k corresponds to an independent tree structure; Ω is a regularization term. For the binary classification which is considered in this study, the logarithmic loss function is used (3.33). In order to learn the decision trees represented with function f_k, it is necessary to define tree parameters, such as a structure of the tree and leaf scores (weights). All trees are built in a stage-wise manner using an additive training strategy by adding one new tree at a time. The prediction value at step *t* is defined as follows:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i). \tag{5.4}$$

At each of the above described steps, we add a new tree that optimises a given objective function. This is achieved using the gradient descent technique. More specifically, gradient boosting of decision trees can be viewed as gradient updates by additional trees learned one at a time. Gradient descent implemented in XGBoost package computes a second-order derivative by applying a second order Taylor approximation (Chen and Guestrin, 2016). This provides a further improvement over the conventional gradient descent technique. The resultant regularised objective at step *t* is defined as:

$$Obj^{(t)} = \sum_{i=1}^{N} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \sum_{i=1}^{t} \Omega(f_i)$$
$$\approx \sum_{i=1}^{N} \left[l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \sum_{i=1}^{t} \Omega(f_i) \quad (5.5)$$
$$\approx \sum_{i=1}^{N} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t)$$

where $g_i = \partial_{\hat{y}_i^{(t-1)}} l(\hat{y}_i^{(t-1)}, y_i)$ and $h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(\hat{y}_i^{(t-1)}, y_i)$ are the first (gradient) and the second (Hessian) order derivatives of the loss function.

Complexity of the decision tree

The bias-variance trade-off is the main problem in supervised ML. Ideally, a classifier should be able to accurately capture the properties in the training data as well as to generalise to the unseen training set. The high complexity of the model may result in a high variance problem or so-called overfitting. In order to decrease the complexity of the model, in our case complexity of the decision tree, regularisation methods are introduced. The complexity of the decision tree can be influenced by a number of parameters, such as the number of terminal nodes in the tree, the depth of the tree and others. Shallow trees are known to have low variance and high bias and are more commonly used for the additive tree models. As the number of trees increases, the bias tends to decrease, while the variance is increasing.

The regularization term $\Omega(f)$ included in the objective function (5.5) is aimed at controlling the complexity of each individual decision tree f and it is defined as follows:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad (5.6)$$

Each f corresponds to an independent tree structure with a vector of leaf weights w on the j^{th} leaf; T is the number of leaves in the tree. The number of terminal nodes is penalized with the γ parameter; weights optimisation is performed using L2 norm, to encourage smaller weighs. The regularization term Ω, helps to smooth the learnt weights to avoid over-fitting.

Regularization

Another regularisation technique of gradient boosting aimed at avoiding overfitting was proposed in (Friedman, 2002). Its main idea lies in introducing randomness in the learning process by using subsampling in each iteration. The XGBoost algorithm includes two types of randomization: row and column subsampling. Row subsampling was shown to improve the

performance of the classifier by using a randomly selected fraction of training examples without replacement. Another way to introduce randomness is achieved with column (feature) subsampling, where at each iteration a decision tree is constructed using a random subset of features. The additional regularization technique is shrinkage (learning rate). It scales newly added weights by a factor, η, after each step of tree boosting. This reduces the influence of each individual tree and leaves space for future trees to improve the model.

The Hessian of the objective function regulates the number of points in the node and represents the level of purity in the node. Setting a minimum allowed number of instances which allows further split in the node can be used to reduce the complexity of the model by penalizing very deep trees.

Learning tree structure

Another important aspect of decision trees is their structure. The tree is defined as follows:

$$f_t = w_{q(x)}, w \in R^T, q: R^d \rightarrow \{1,2,..,T\} \tag{5.7}$$

where the mapping q is a function which assigns each data point to the corresponding leaf, T is a number of leaves, and w is a vector of weights on the leaves. According to (5.5) the resultant regularised objective at step t is defined as:

$$\begin{aligned} Obj^{(t)} &\approx \sum_{i=1}^{N}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2 = \\ &= \sum_{j=1}^{T}\left[w_j \sum_{i\in I_j} g_i + \frac{1}{2}w_j^2\left(\sum_{i\in I_j} h_i + \lambda\right)\right] + \gamma T \end{aligned} \tag{5.8}$$

where $I_j = \{i|q(x_i) = j\}$ defines a set of indices of data points that are assigned to the j-th leaf using function $q(x)$, with q as a tree structure. An optimal weight, w_j, for region j is defined as follows:

$$w_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda} \tag{5.9}$$

The corresponding objective can be rewritten as:

$$Obj^* = -\frac{1}{2}\sum_{j=1}^{T}\frac{(\sum_{i\in I_j} g_i)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T \tag{5.10}$$

Learning the tree structure of f_t implies deciding on how to split features. After each split, the leaf is converted to an internal node and new left (I_L) and right (I_R) nodes are generated, $I = I_L \cup I_R$. The gain of each split is defined as follows:

$$Gain = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma \quad (5.11)$$

The first and second terms in the formula represent scores for the left and right leaves added, the third term is a score of the original leaf prior to the split. The tree is built to the predefined maximum depth. The nodes with negative gain are then pruned in a bottom-up order. It can be seen that if the gain is smaller than the regularization parameter γ, the split is not added. A general overview of the boosting algorithm is presented below:

Algorithm 5.1: Gradient Boosting

Input:

Dataset *D*

A loss function *l*.

A base learner $\mathcal{L}_\Phi$, with base models $\phi_1, \dots, \phi_k \in \Phi$.

The number of iterations *K*.

The learning rate η.

1 Initialize $\hat{y}^{(0)}(x) = \hat{f}_o(x) = \hat{\theta}_0 = \arg\min \sum_{i=1}^{N} l(y_i, \theta)$

2 **for** $k = 1,2, \dots, K$ **do**

3 $\hat{g}_k(x_i) = \partial_{\hat{y}_i^{(k-1)}(x)} l\left(y_i, \hat{y}_i^{(k-1)}(x)\right)$;

4 $\hat{h}_k(x_i) = \partial^2_{\hat{y}^{(k-1)}(x)} l\left(y_i, \hat{y}_i^{(k-1)}(x)\right)$;

5 $\hat{\phi}_k = \arg\min_{\phi\in\Phi} \sum_{i=1}^{N} \frac{1}{2}\hat{h}_k(x_i)\left[\left(-\frac{\hat{g}_k(x_i)}{\hat{h}_k(x_i)}\right) - \phi(x_i)\right]^2$;

6 $\hat{f}_k(x) = \eta\hat{\phi}_k(x)$;

7 $\hat{y}^{(k)}(x) = \hat{y}^{(k-1)}(x) + \hat{f}_k(x)$;

8 **end**

Output:

$\hat{y}(x) = \hat{y}^{(K)}(x) = \sum_{k=1}^{K} \hat{f}_k(x)$

An example of an individual decision tree constructed using HRV features for the prediction of the health status of the preterm is represented in Figure 5.15. As mentioned earlier, a decision tree can use missing data. For a given tree example, the missing values have the same direction as values that follow the 'yes' decision.

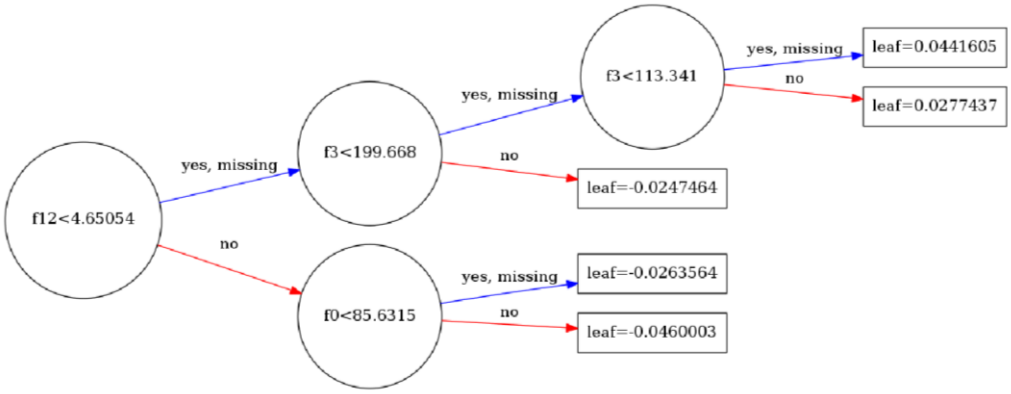


Figure 5.15: The visualization of an individual decision tree constructed using XGBoost; f12, f0, and f3 correspond to the HRV features: RMSSD, VLF, and LF/HF respectively. Every decision tree is trained to find out whether preterm has a good or poor outcome. Prediction score is assigned to every leaf – prediction for data points which fall into the given leaf. The final prediction is then obtained as the sum of scores predicted by each of the trees.

Feature importance

The previously described *gain* parameter allows for an estimation of the importance of each feature and their subsequent ranking. The importance is represented as an improvement in accuracy which is attributable to a given feature in all branches of the decision tree ensemble. The calculation of the gain is performed for every feature in each node of the tree. In this study, we have also investigated the interaction of various features (Far0n, 2018). The interaction between two features is defined as a path of length two within a tree and is calculated by summing the node gains (Eq.(5.11)) along the corresponding decision path. The process is repeated for every pair of features in each decision tree from the tree ensemble.

5.6.3 Model selection and performance assessment

Whereas the predictive power of each feature is computed on the whole dataset for descriptive statistics, the application of ML tools requires clearly defined model selection and performance assessment routines. The leave-one-out (LOO) subject independent performance assessment is used in this work to estimate the generalisation error. All but one subject's data are used for training and the remaining subject's data are used for testing. The procedure is repeated until each patient has been a test subject (Figure 5.16). The LOO method is known to be an almost unbiased estimation of the true generalization error (Vapnik, 2006), i.e. the error reported with this routine is the most correct estimation of the error this system would get by testing on a separate unseen dataset of infinite size once it is trained on all available data.

The regularised booster classifier used in this study has a number of user-tunable parameters (hyperparameters). The most important ones are the depth of the tree, the feature and data

subsampling ratios, as well as the number of iterations which are a function of the learning rate η.

In order to balance between underfitting and overfitting, an optimisation of hyperparameters is performed using an exhaustive search through the manually specified range for each hyperparameter. The hyperparameters were tuned using the stratified 5 times 2-fold cross validation (CV) (Dietterich, 1998), performed on the training data. This allows for a balanced representation of both classes in each fold. The CV folds were designed in a way that preserves the subject integrity – that is no subject data appear in two different folds. This allows for a model selection routine that has the optimization criteria that matches the one of the LOO performance assessment routine – subject-independent evaluation. In other words, the model selection routine searches for hyperparameters that maximise the performance on unseen patients in the internal CV, and these parameters are then used to train the model which is assessed on a single unseen patient in the external LOO loop. It can be seen that the performance assessment routine is independent of the model selection routine, and the testing patient is not seen or used for training the classifier or tuning of other system parameters at any time. The best performing parameter set is saved for each LOO iteration. The analysis of stability of the classifier is then performed by examining the most frequent set of hyperparameters used to train the classifier.

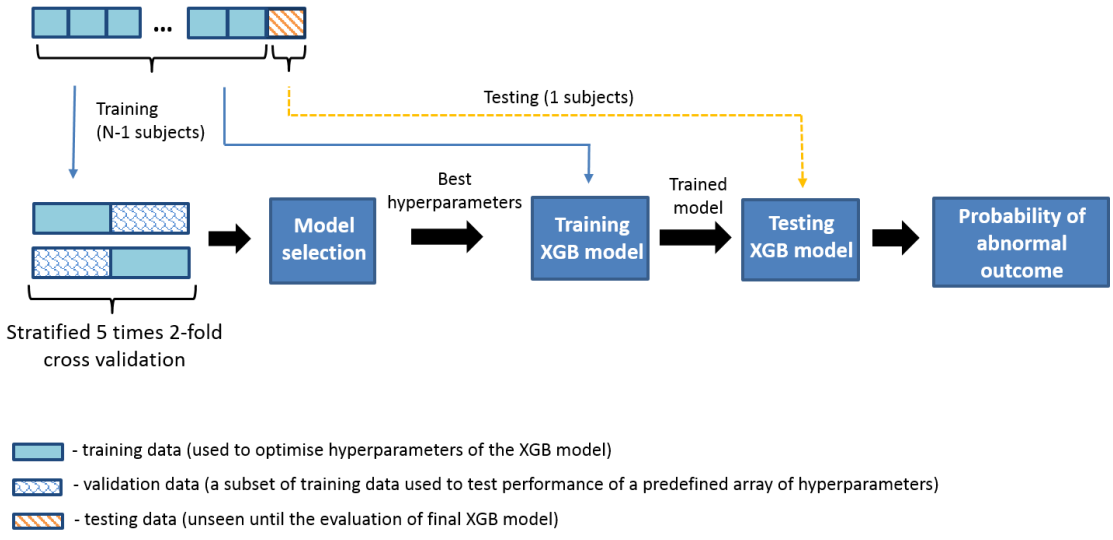


Figure 5.16: A diagram of LOO subject independent performance assessment and 5 times 2-fold CV model selection routines.

The LOO performance was similarly measured by the AUC metric and computed on the epoch and subject levels. When an epoch of the physiological data (a vector of EEG or HRV features) is fed into the designed XGB classifier, the probability of an abnormal outcome is returned for this epoch. These probabilities are then compared to the ground truth (outcome scores) and the AUC is computed across all epochs. The ground truth for each epoch is taken as the ground

truth for the whole patient. The mean probability across all epochs for a given patient is also calculated, resulting in a single probabilistic value per subject which represents the level of the algorithmic support for the abnormal outcome decision. The subject level AUC is then obtained by contrasting the averaged probabilities with the outcome labels. The two ways to compute the AUC are designed to assess the instantaneous accuracy of the classifier (epoch-based AUC) as well as the accuracy when the whole recording is available (subject-based AUC).

5.6.4 Statistical inference and out-of-sample predictive modelling

The discriminative power measured by the AUC represents a statistical descriptive linear inference of a separate signal characteristic. This measure of separation is computed from all the data and small intra-subject data variability may impact the conclusions regarding whether the observed phenomenon generalizes across various subjects. To assess the out-of-sample predictive power of a feature or combination of features the subject independent performance assessment must be used. With ML methods, the LOO performance assessment allows 1) to estimate whether features bear predictive power that generalizes across patients, 2) to estimate the predictive power of nonlinear correlation between the feature and the label, 3) to estimate the above for a combination of features. Figure 5.7 indicates that a combination of features can indeed increase the separation between good and poor outcomes. The challenging problem of statistical multivariate analysis can be addressed with ML.

5.6.5 Combination of features with boosted decision trees

The aim of any decision support system should be to provide high sensitivity, correctly detecting as many (ideally all) of the 'unhealthy' preterm babies in the NICU. So there is a trade-off here – the choice of a tighter threshold (e.g. MAP $\leq$ GA) will provide a better performance. This performance, however, will come at the cost that some unhealthy babies, who do not exhibit any such deep dips in BP across their recording, will be excluded totally from the analysis. This would lead to the risk that a substantial number of unhealthy neonates (in our case 4) would not be processed and correctly detected by the proposed decision support system.

The combination of features with boosted decision trees was investigated for Set 3 ($MAP \leq GA+4$) since for this threshold the separation between healthy and unhealthy newborns is more challenging while all the data can be used. In order to predict the short-term outcome, the HRV and EEG features were separately fed into boosted decision tree classifiers. The results of the performance of classifier-based systems are presented in Table 5.5. The prediction systems were designed to run on **All epochs** and **Hypotensive events** and were scored with respect to their epoch or subject level accuracies.

Table 5.5: AUC for short-term outcome prediction using various combination of HRV, EEG and BP features (MAP≤GA+4). AUCs of the best performing HRV- and EEG-based systems are in bold.

	Predictions on **Hypotensive events**			
	HRV	HRV & MAP	EEG	EEG & MAP
	23 subjects, 1488 epochs		25 subjects, 2564 epochs	
Epoch AUC	**0.92**	0.91	0.73	0.72
Subject AUC	**0.97**	0.96	0.76	0.81
	Predictions on **All epochs**			
	HRV	HRV & MAP	EEG	EEG & MAP
	23 subjects, 6217 epochs		25 subjects, 7030 epochs	
Epoch AUC	0.83	0.88	0.62	**0.74**
Subject AUC	0.94	0.95	0.73	**0.85**

It can be seen that feeding all thirteen HRV features extracted only from **Hypotensive events** achieves an AUC of 0.92 and 0.97 for epoch and subject level scoring, respectively. EEG features reach an AUC of 0.73 and 0.76, respectively.

The HRV-based system which is designed to operate on **All epochs** results in an AUC of 0.83 and 0.94, for the epoch and subject level metrics. The performance of the EEG-based system is 0.62 and 0.73, respectively.

Focusing on the best performing HRV system for **Hypotensive events**, Figure 5.17 shows that the ranking of the various HRV features is consistent with statistical descriptive analysis (Table 5.3). The importance is computed as an average of 23 iterations of the patient independent LOO procedure. It can be seen that RMSSD, MeanRR, LF/HF, ApEn and LF are the top five most important HRV features for short-term health outcome prediction. Figure 5.18 illustrates the top ten two-feature interactions which have the greatest contribution to the preterm outcome prediction. It can be seen that the RMSSD feature is involved in many of the important interactions. This result is consistent with the single feature statistics, where RMSSD showed the highest predictive power (AUC=0.87, Table 5.3, Set 3). At the same time, MeanRR is the feature which is involved in the majority (7 out of 10) of the top ten two-way interactions. This indicates that the absolute HR information is complementary to many HRV features for the chosen task and HRV must be considered in the context of mean HR.

Figure 5.19 and Figure 5.20 represent the ranking of the EEG and BP features for the best EEG-based classifier (all epoch, EEG and MAP features). The most important feature is RP in 3-8 Hz EEG sub-band. Unlike the HRV-based system, where its statistical inference is comparable with the out-of-sample modelling with the best RMSSD feature, for the EEG-

based system RP (3-8 Hz) accounts for the highest classification gain, while its AUC (based on 7030 epoch – all available data) is only equal to 0.59.

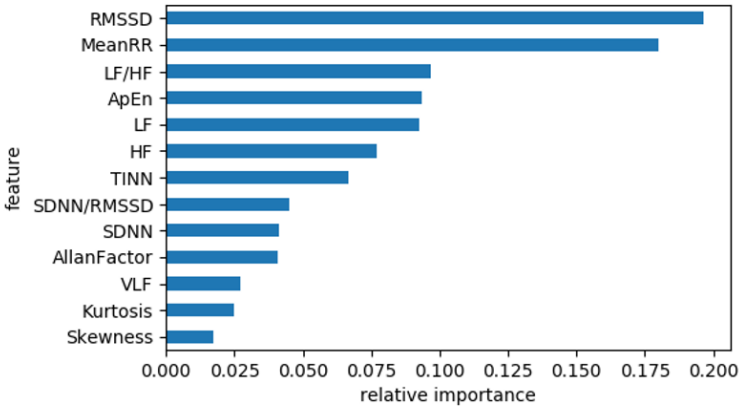


Figure 5.17: Mean of the feature importance (gain) reported by the boosted decision tree classifier trained on 13 HRV features extracted during episodes of low BP (MAP≤GA+4).

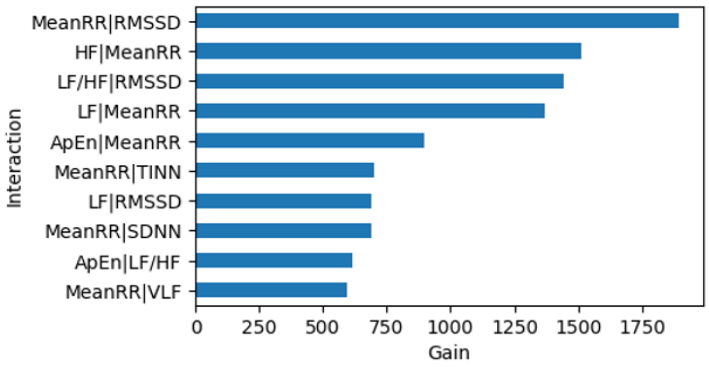


Figure 5.18: Representation of the importance (gain) of the top ten two-feature interactions for the short-term outcome prediction. The system is trained on the 13 HRV features extracted during the episodes of low BP (MAP≤GA+4).

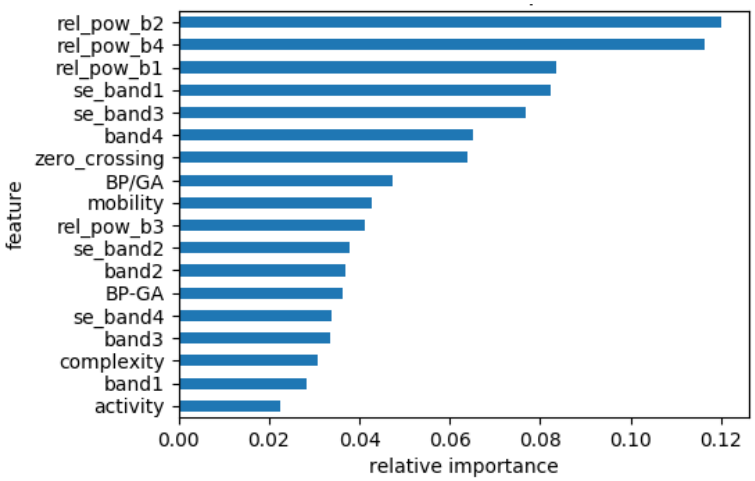


Figure 5.19: Mean of the feature importance (gain) reported by the boosted decision tree classifier trained on all epoch of the EEG and BP features.

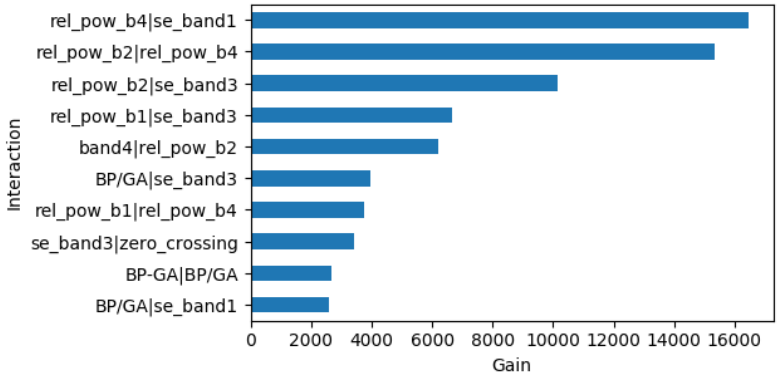


Figure 5.20: Representation of the importance (gain) of the top ten two-feature interactions for the short-term outcome prediction. The system is trained on all EEG and BP features extracted from the full recordings.

Stability of the classifier

Figure 5.21 illustrates the most frequently selected sets of hyperparameters during the internal CV for each of the iterations of the LOO routine, for the best performing classifier based on the HRV features extracted during **Hypotensive events**. The radius of each sphere indicates the frequency with which a particular set of hyperparameters was selected.

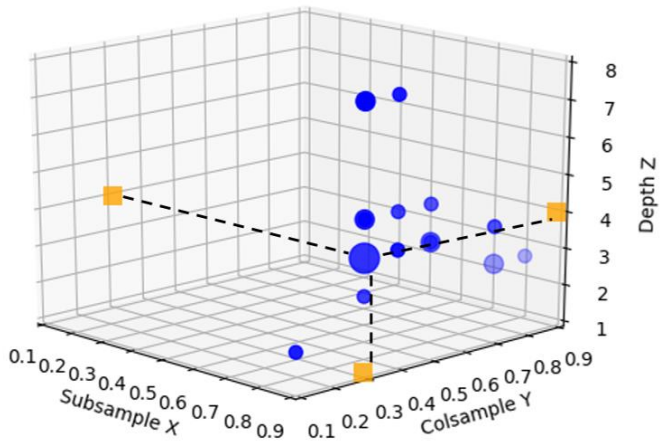


Figure 5.21: The density of selected tensors of three main hyperparameters obtained during the LOO routine for the HRV-based systems for Hypotensive events (MAP≤GA+4). The most frequently selected parameters are Subsample=0.9, Colsample=0.3 and Depth=4. The projections of the parameters are represented with dash lines.

It can be seen that while the exhaustive search was performed on a wide range of hyperparameters, the choice of the best tensor of hyperparameters is stable as represented by the densely located sets of best hyperparameters. According to the CV performance metric, the best set of hyperparameters for the classifier trained on HRV features from Hypotensive events dataset is located at (0.9, 0.3, 4), where x, y, and z are the fraction of the data, features and depth of the tree, respectively. The tree depth of 4 corresponds to a tree, which is capable of creating complex decision boundaries and benefiting from the interaction between features.

Every tree is built on a randomly selected set of 90% of data and 30% of all features. The fact that 30% of features are selected supports the conclusions from the analysis of feature importance and the statistical analysis which indicates the presence of many irrelevant features. If we were to generate a single model on all available training data to be used in clinical practice, then these hyperparameters would be used to train the final classifier.

5.7 Feature selection

The performance of the data-driven algorithms is usually highly dependent on the feature set. Therefore it is crucial to use features which are relevant for a given task. There exist different feature selection techniques, which allow retaining only the most useful variables in order to improve the performance of the classifier as well as to reduce the computational time (Guyon and Elisseeff, 2003). The main feature selection methods are categorised as 1) filter techniques; 2) wrapper methods; and 3) embedded techniques (Figure 5.22).

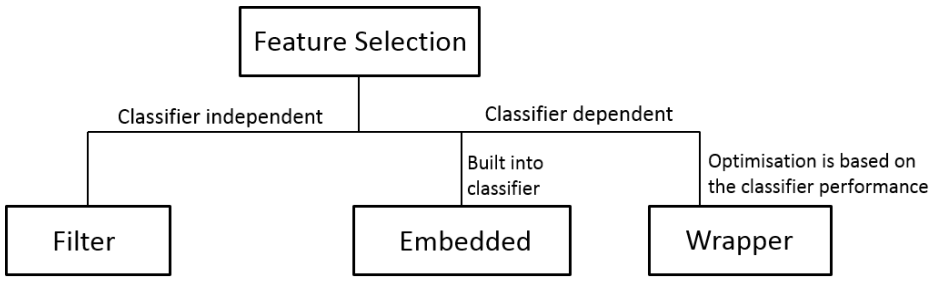


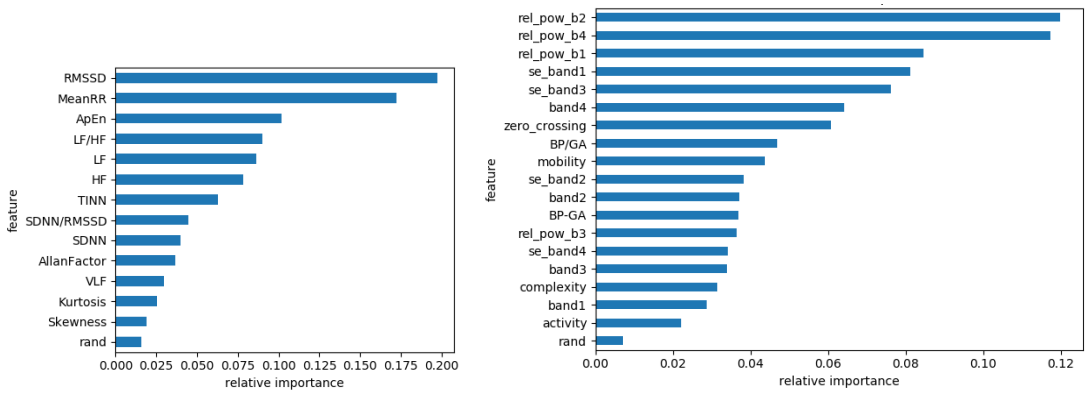
Figure 5.22: Feature selection methods.

The simplest feature selection method is a filtering technique. This method is independent of the classifier and is based on a specified metric, which defines the relevancy of the feature and the labels. From Table 5.3 and Table 5.4 it is clear that according to simple statistical inference (based on AUC), both EEG and HRV features bear a certain predictive power, which is different to random chance, and therefore may be retained.

Random feature

When working with boosted decision trees, the practice of adding a redundant feature to the original feature set is commonly used. This is done in order to ensure that the feature sets do not contain any unrelated variables which are not useful for the out-of-sample predictive modelling. The random feature is generated by random sampling from a uniform distribution. If all features are relevant for the out-of-sample predictive modelling, the random feature is expected to appear on the bottom of the feature importance list. If this is not the case, all features with the gain lower than the gain of the redundant feature should be discarded. This can potentially help to further improve the performance of the classifier.

In this work, a randomly generated feature was injected into the best HRV- and EEG-based classifiers. From Figure 5.23 it can be seen that the redundant feature has the lowest gain for both classifiers. When introducing random feature (noise) into the system, a slight drop in the performance of the HRV classifier (hypotensive events) is observed: epoch level AUC=0.91 vs AUC=0.92, subject level AUC=0.95 vs 0.97. A similar result was obtained for the EEG-based classifier (all epoch, EEG and MAP features), where the randomly generated feature showed the lowest gain, while the performance of the classifier was not affected. Obtained results indicate that all EEG and HRV features are useful for the out-of-sample predictive modelling.



(a) (b)

Figure 5.23: Mean of the feature importance (gain) reported by two boosted decision tree classifiers trained on the (a) HRV features extracted during episodes of low BP (MAP≤GA+4), and (b) EEG and MAP features extracted from all epoch. An additional randomly generated feature 'rand' is incorporated. The obtained results indicate that all EEG and HRV features bear some predictive information when compared to the random one.

Top features

XGBoost, as well as other tree-based classifiers, can be considered as embedded feature selection techniques. This is due to the fact that each feature is evaluated as a potential splitting point. Consequently, variables which are not able to discriminate between classes will not be included in the decision tree. This property along with the regularization techniques made XGBoost a widely applied feature selection tool in Kaggle ML and data mining challenges. More specifically, instead of using the full feature set, even though the performance of each feature is better than the random one, only a certain number of features with the highest gain are used. Features with a high gain are known to be the most important for the final outcome prediction and therefore are usually incorporated in high-gain feature interactions.

In this study, in order to further boost the performance, the top features of the EEG- and HRV-based classifiers were used. The inclusion criteria was set to gain higher than 0.05. This resulted in the top seven features for HRV-based classifier (RMSSD, MeanRR, ApEn, LF/HF,

LF, HF, and TINN), and EEG-based classifier (RP (0.3-3 Hz), RP (3-8 Hz), RP (15-30 Hz), SE (0.3-3 Hz), SE (8-15 Hz), SE (15-30 Hz) and zero crossing feature). The obtained results indicate that the AUC of the HRV system instead of improving has actually decreased (AUC=0.92 vs AUC=0.89 and AUC=0.97 vs AUC=0.95 for the subject- and epoch-level performances respectively). At the same time, the reduced feature set of EEG classifier did not affect the classifier performance.

Forward feature selection

Wrapper feature selection techniques are based on the performance of the classier and yield classifier specific sub-set of features. The main advantage of this method over the filter feature selection is that the wrapper technique is directly targeted at minimisation of the classification error for a given training data set. There are different ways to find the optimal feature subset. An exhaustive search technique, which evaluates each possible combination of feature, is usually not a valid option. This approach is computationally expensive, especially if the dataset is comprised of a high number of features.

In this work, we have incorporated a sequential forward search. This algorithm expands an empty feature subset by adding one new feature at a time based on its CV performance. A schematic representation of the forward feature search wrapped into the subject independent LOO performance assessment is represented in Figure 5.24. It can be seen that prior to the selection of hyperparameters, the feature selection step is performed. This step results in a subset of '*selected features*'. More specifically, on the first stage, the CV is performed for each feature separately. A feature which resulted in the highest CV AUC is then added into the '*selected features*' subset. On the next stage, the CV is performed for a number of combinations of the best feature (from the '*selected features*' subset) with other features (which have not been included into the subset). A new feature which has further improved the CV AUC is also added into the '*selected features*' subset. This procedure repeats (adding a single feature at a time) until the CV performance cannot be further improved. As a result, each LOO interaction may contain different subsets of features.

The performed forward feature search for the HRV and EEG classifiers did not allow to further improve the classification performance: the HRV system resulted in AUC=0.95 and AUC=0.88 for subject and epoch levels, and in AUC=0.82 and AUC=0.75 for the EEG system respectively.

The construction of the decision trees incorporates the inclusion of relevant features only based on a certain metric (gain in this study), which measures the improvement achieved by adding a certain feature into the tree. In our study, the HRV and EEG classifiers incorporate a

relatively small number of features. This may suggest that the extracted features are complementary to each other and that the feature selection procedure could have been more effective for a larger feature set. At the same time, it is also important to acknowledge the limitations of the forward feature selection technique. More specifically, this method does not allow to remove features from the '*selected features*' subset and therefore, it fails to investigate other possible combinations of features, which could have resulted in better performance.

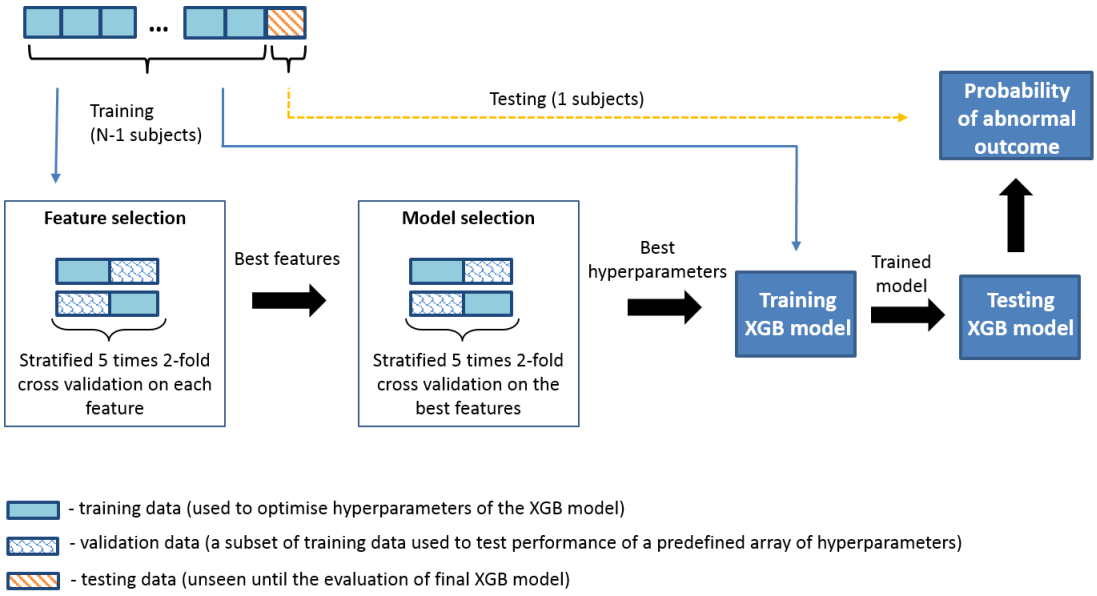


Figure 5.24: A diagram of LOO subject independent performance assessment and 5 times 2-fold CV feature and model selection routines.

5.8 Decision support tool

The designed systems allow for the continuous observation of the probabilities of morbidity for the preterm infant. Figure 5.25 depicts the probabilistic traces for two subjects, one with a healthy and one with an unhealthy outcome. Figure 5.26 presents the probabilistic output of the same unhealthy subject as in Figure 5.25 for the **Hypotensive events** and **All epochs** systems along with the trace of MAP to better illustrate the purpose and the functionality of each system.

The GA-based rule on the definition of hypotension is widely used in the clinical practice for the treatment of low BP in the preterm to decide whether current low BP is of any risk to the wellbeing of the neonate. In this study, the HRV-based systems, **All epoch** and **Hypotensive events,** were scored with respect to their epoch and subject level accuracies. Each system continuously outputs a probability of morbidity for every five-minute window in real-time as shown in Figure 5.26. Such systems can be used in clinical practice as a decision support tool for monitoring of preterm neonates, who may have low BP. The MAP recording along with

the probability traces of two systems, **All epochs** and **Hypotensive events**, give additional insight into the interrelation between HRV, BP and neonatal health outcome in the context of episodes of low MAP. It can be seen that whereas system trained on **All epochs** (Figure 5.26 (b)) outputs probabilities for every segment of data, the system trained on **Hypotensive events** (Figure 5.26 (c)) outputs the probabilities only on the segments where MAP falls below the predefined threshold. The system designed to operate on **Hypotensive events** outperforms the one designed on **All epochs** as shown in Table 5.5. Similarly, in Figure 5.26, the continuous and cumulative average probability of morbidity for a given unhealthy neonate is higher using **Hypotensive events**, 0.95, as compared to 0.85 for **All epochs**.

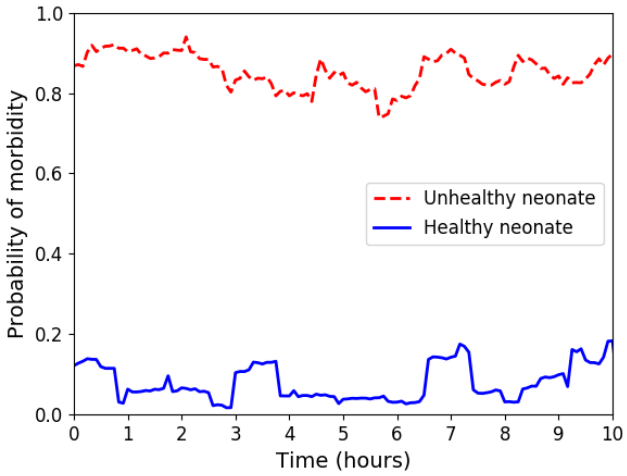


Figure 5.25: An example of the system output as a continuous probabilistic trace obtained during 10 hours for one healthy (GA=28 weeks, blue solid line) and one unhealthy (GA=23 weeks, red dashed line) patients. The system is trained and evaluated on the **All epochs** dataset.

Non-invasive techniques for BP measurement are known to be unreliable in small and sick infants. An excessive intervention in sick preterm infants is undesirable, as a result, a gold standard invasive BP recording at times may be difficult to obtain (Weindling, 1989). On the other hand, non-invasive ECG is routinely recorded in preterms. It is possible to identify 3 decision support tools: 1) outcome predictor based on HRV, 2) outcome predictor based on HRV and MAP, and 3) outcome predictor based on HRV and MAP for use during episodes of hypotension. It can be seen from Figure 5.26 that the cumulative average of probability stabilises quickly and thus can approximate the final average probability which is used in the subject-level assessment of accuracy. As such with the cumulative probabilistic output the discriminative capacity of the system moves gradually from epoch-level AUC to subject-level AUC as a function of monitoring time. Table 5.5 shows that in the absence of BP registration (**Hypotensive events** results thus are automatically excluded as they require MAP to identify the hypotensive episodes), the decision support tool (1) that uses only HRV features can reach an AUC between 0.83 - 0.94 (the range between the epoch-level and subject level accuracies). This is represented in Figure 5.26 (b) as a red solid thin line for instantaneous predictions and

green dashed bold line for cumulative average probabilistic values. The decision support that uses HRV and MAP but operates on every epoch (2) obtains an AUC between 0.88 and 0.95. The best performing system (3) which additionally requires MAP values to be below a predefined threshold achieves an AUC between 0.92 - 0.97. This is represented in Figure 5.26 (c) as a solid blue thin line for instantaneous predictions and solid orange bold line for cumulative average probabilistic values. This indicates the accurate decision support tools can be designed depending on the registered signals and the presence of hypotensive events.

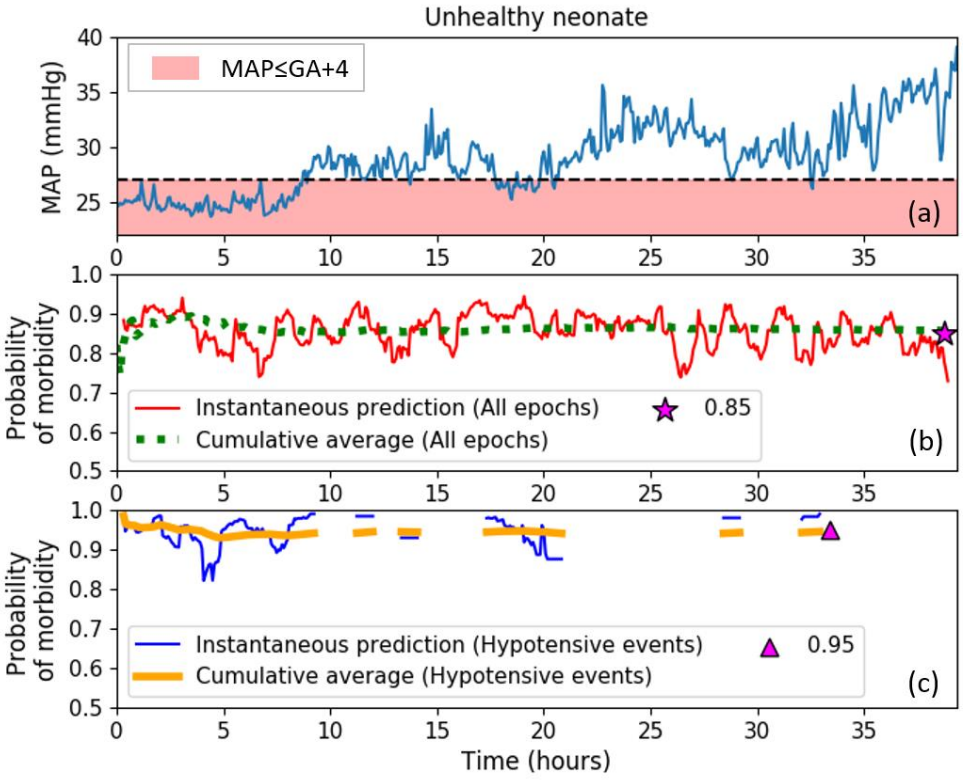


Figure 5.26: Comparison of the probabilistic traces for an unhealthy neonate (GA=23 weeks) obtained from the two models trained on HRV features extracted from either the All epochs (b) or Hypotensive events (MAP<GA+4) (c) datasets. The model trained on the All epochs (b) is represented by instantaneous (red solid thin line) and cumulative (green dashed bold line) probabilistic values. The Hypotensive events model (c) is represented by the instantaneous probabilistic values (solid blue thin line) and the cumulative average of prediction (solid orange bold line). An average of the morbidity prediction for both models is 0.85 and 0.95 correspondingly.

5.9 Discussion

Combination of features with boosted decision trees

It can be seen from Table 5.5 that the HRV-based classifier showed higher discrimination capacity than that of the EEG-based classifier. This result was expected, as from a clinical point of view HRV is seen to be more indicative of the current physiological stability of the neonate. It can be observed that the subject-level performance is higher than the epoch-level. The accumulation of probabilistic information across the whole recording increases the

discriminatory capacity of the classifier. The performance reported at the subject level assumes the availability of the whole recording before decision making. This means that the system achieves this performance for prognostication purposes rather than for on-the-fly monitoring.

Similarly, the performance for **Hypotensive events** is higher than that for **All epochs** both for HRV and EEG-based classifiers. For example, the HRV-based system improves from an AUC of 0.83 to an AUC of 0.92 when observing only **Hypotensive events.** An additional experiment was conducted where MAP information was provided to the classifiers in the form of two additional features: *MAP–GA* and *MAP/GA*, which are normalised by *GA*. This additional information has improved results on **All epochs** for both HRV and EEG-based systems. More specifically, for the model trained on the HRV & MAP features, the AUC increased from 0.83 to 0.88. Similar behaviour can be observed for the classifier trained on the EEG & MAP features, where the AUC has increased from 0.62 to 0.74 (Table 5.5). No consistent benefit was observed for **Hypotensive events** where the dataset has been already pre-selected based on the MAP thresholds and thus MAP features do not bring any new information. The overall best results both for the epoch and subject-level metrics are obtained with the HRV-based system on **Hypotensive events** with an AUC of 0.92 and 0.97, respectively.

Limitations

The study has a number of limitations. Prior to feature extraction, artefacts were removed from the EEG, ECG and MAP signals. Different segments of data were corrupted by ECG and EEG artefacts which resulted in different data content for HRV and EEG. The combination of HRV and EEG features was thus only possible on a dramatically reduced amount of data where both EEG and ECG were simultaneously free of artefacts and resulted in relatively poor performance due to data scarcity and is not presented in this study.

The definitions used to define low BP are arbitrary. The first definition we chose was based on the GA rule, which seems to be one of the most commonly utilized definitions. The other two definitions /thresholds used were arbitrary and are not supported by any definitions in clinical use. We chose these because the natural rise in BP over the first day of life is approximately 5mmHg. These two definitions of +2 and +4 would remain less than the normal rise noted in healthy preterm infants. Other limitations include the potential effect of the medications, intubation or mechanical ventilation on BP.

In this study, EEG was recorded from preterms over a wide range of GAs – from a minimum GA of 23 weeks to a maximum GA = 31 weeks. However, it is known that an accurate interpretation of the premature brain requires an appreciation of the maturational features for

neonates of all GAs. This is due to the fact that the baseline pattern of the EEG changes with maturation (Pavlidis et al., 2017). Preterm EEG is characterized by discontinuous activity, where high voltage delta activity (bursts) is followed by low voltage IBI. The duration of discontinuity decreases with increasing GA. More specifically, between 23 and 27 weeks of GA the duration of IBI is <60 seconds and for GA between 30 and 31 weeks, IBI decreases to ≤20 seconds. The morphology of busts also changes with GA. Increasing GA is characterized by decreased amplitude in delta and theta bands and appearing of superimposed fast activity (Jennekens et al., 2012; Pavlidis et al., 2017). Therefore, analysis of EEG in the first few days in preterm neonates with varying GAs is another potential reason for poor sensitivity of EEG to the health status of the preterm neonate.

5.10 Conclusion

The results of this chapter indicate that HRV is sensitive to BP and the features extracted during the episodes of low MAP improve the prediction of the short-term outcome for the preterm neonates. HRV features extracted during episodes of low BP were shown to correlate with short term health outcome represented by clinical course scores of preterm neonates, with a single best feature (RMSSD) reaching an AUC of 0.87. EEG features were, as expected, shown to be less predictive of short-term health outcomes, when compared to HRV, with the best performing feature (SE) yielding an AUC of 0.75.

Combining multimodal data, HRV, EEG, and BP, an objective decision support tool for clinical prediction of short-term outcome in preterms can be constructed. ECG with or without BP records are usually available soon after birth and this work presents a promising step towards the use of multimodal data in building an objective decision support tool for the clinical prediction of short-term outcome in preterms. In the long run, this tool can potentially help clinicians to properly select patient-specific treatment procedures if required.

Chapter 6: Automated assessment of 2-years neurodevelopmental outcome using early EEG recordings

This chapter presents a study towards the development of an automated system for the long-term neurodevelopmental outcome assessment for preterm neonates. Here two approaches are proposed, the first one incorporates the classical feature-based approach followed by a classifier; the second system performs end-to-end learning, by applying convolutional neural networks to the raw EEG signals. Results of both systems are presented and contrasted.

6.1 Clinical problem and motivation

A recent study by Pierrat et al., (2017) reported that over the past two decades the survival rate of the preterm infants has improved. These neonates, however, remain at a high risk of neurodevelopmental delay. The lack of proper development and maturation of the brain and other vital organs has resulted in impaired speech and language outcomes for the very preterm infants with lasting effects in childhood and adolescence (Vohr, 2014). Preterms were shown to be at a high risk of social and emotional disorders such as depression and anxiety as well as psychiatric disorders (Botting et al., 1997). Overall, these infants exhibited significantly higher disorganized behaviour (e.g. motoric, attentional, reactivity) as compared to full-term neonates (Als et al., 1988). Early interventions for preterm neonates have demonstrated a positive influence on their motor and cognitive development (Spittle et al., 2015). This may suggest that a necessary treatment initiated soon after birth can potentially improve the outcome of the preterm neonate.

The long-term prognostic capability of the EEG was reported in a number of previous studies (Lloyd et al., 2016; Pressler et al., 2001). A meta-analysis by Sinclair et al., (1999) showed that slow activity, burst suppression, and low voltage are indicative of an increased risk of

neurodevelopmental handicap. Diffusion tensor imaging (DTI), which allows for the analysis of the white matter microstructure in 3D space is yet another method of brain assessment which has shown to be predictive of the neurodevelopmental outcome for neonates who suffered from perinatal arterial ischemic stroke (van der Aa Niek E. et al., 2011). A study by Roze et al., (2015), which also incorporated DTI obtained from the preterm infants with periventricular hemorrhagic infarction (bleeding into the brain), has reported an association between DTI and motor outcome of infants assessed at 15 months.

In the recent study by Pavlidis et al., (2019) a standardised assessment of the preterm EEG was proposed. Two experts graded the two hours of EEG into 4 categories: mild, moderate, severe and markedly abnormal EEG. The developed grading system is based on the normal and abnormal EEG waves which were visually estimated while accounting for the GA and information about drugs administered. Within the scope of the PhD thesis of Lloyd (2019) the statistical association between the extracted EEG grades and the 2-year neurodevelopmental outcome (quantified using the Bayley scales of infant development III) has been investigated. The obtained findings have demonstrated promising results of the predictive capability of the EEG with respect to the long-term neurodevelopmental outcome of the preterm neonate.

The procedure of manual EEG grading (Pavlidis et al., 2019) is challenging as it is a time-consuming process which requires domain-specific knowledge. Nowadays, computer-based analysis of signals is widely applied for healthcare applications (Faust et al., 2012). In this context, the main objective of this study is to investigate the predictive capability of the early EEG recorded at discharge from the NICU with respect to the 2-year neurodevelopmental outcome using ML techniques. This can potentially contribute to the faster, objective and more efficient identification of at-risk infants.

The assessment of the long-term outcome based on the early EEG is a challenging task. Unlike the studies conducted in Chapter 4 and Chapter 5, where the outcome was assessed within the first hours (CRIB) and days (CCS) of life, the gap between the time of the EEG recording (35 weeks GA) and the neurodevelopmental assessment at 2 years of age is much larger (Figure 6.1). During this period of two years, a neonate can be exposed to various uncontrolled risks and possible confounding factors which are not taken into account but can potentially affect the neurodevelopmental process. Acknowledgment of the challenges and limitation of the given problem is important as it will allow for an objective analysis of the obtained results.

6.2 Dataset

The dataset used in this work is a subsample of a larger dataset used by Lloyd, (2019). The EEG recordings from 37 preterm neonates obtained at the NICU of Cork University maternity hospital, Ireland were used in this study. Multi-channel EEG was recorded using a Natus NicOne video EEG machine (CareFusion Co., San Diego, USA) using the international 10-20 system of electrode placement, adjusted for neonates, with the analysis performed on the 8 bipolar channels: F4–C4, C4–O2, F3–C3, C3–O1, T4–C4, C4–Cz, Cz–C3, and C3–T3. EEG signals were sampled at 256 Hz (34 subjects) and 200 Hz (3 subjects). The data were collected at 35 weeks GA, before the discharge of the neonate from the NICU, as shown in Figure 6.1. The duration of recordings from 37 preterms totals 4091 mins (median 120 mins).

For this study, following the approach taken by Lloyd, (2019), the long-term (2-year) neurodevelopmental outcome of the preterm neonate is quantified using the Bayley Scales of Infant Development (Bayley-III) (Bayley, 2006). The details of the Bayley assessment for each subject are provided in Table 6.1. A poor neurodevelopmental outcome was defined using either of two criteria: 1) *composite scales outcome*, defined as abnormal if the value any of the three Bayley subscales (cognitive, language or motor) was less than 85 (Schumacher et al., 2013); 2) *language scale outcome*, defined as abnormal if the value of the language scale alone was less than 85. This resulted in 9 and 7 neonates with the poor long-term outcome for the composite and language scales, respectively. Due to the small number of scores corresponding to poor outcome (<85), neither cognitive nor motor scales were considered independently (4 and 2 for cognitive and motor scales respectively). The schematic representation of the timing of EEG recordings and the long-term outcome assessment is shown in Figure 6.1.

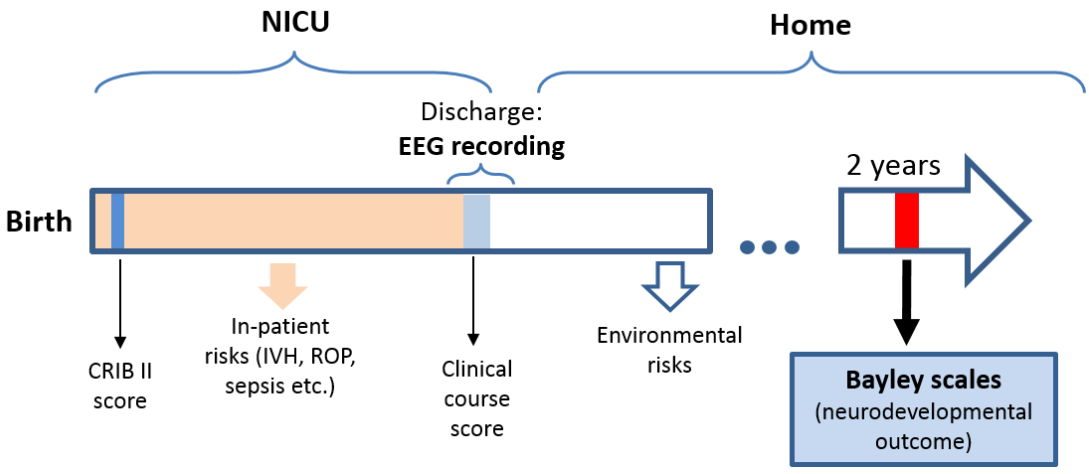


Figure 6.1: The timing of the Bayley scales along with other health scores assigned to an infant in the neonatal intensive care unit (NICU). It can be seen that from the time of the EEG recording (35 weeks GA) to the neurodevelopmental assessment (2 years of age) the neonate is exposed to various risks and confounding factors which are not taken into account but could potentially affect the developmental process.

Table 6.1: Details of the Bayley-III assessment and a corresponding binary outcome. Sick: 24%, healthy: 76% for composite scale; sick: 19%, healthy: 81% for the language scale. The outcome is defined as abnormal (in bold) if 1) the value any of the three Bayley subscales is less than 85 for the composite outcome; 2) abnormal language outcome if the value of the language scale alone is less than 85.

Subject number	Cognitive scale	Language scale	Motor scale	Composite outcome	Language outcome
1	100	94	110	0	0
2	100	91	97	0	0
3	100	91	91	0	0
4	85	**83**	**79**	**1**	**1**
5	95	118	94	0	0
6	100	**79**	91	**1**	**1**
7	115	115	97	0	0
8	95	89	88	0	0
9	105	112	94	0	0
10	95	**77**	103	**1**	**1**
11	**80**	**71**	-	**1**	**1**
12	105	115	97	0	0
13	100	141	100	0	0
14	110	121	100	0	0
15	**70**	**50**	**64**	**1**	**1**
16	100	**83**	100	**1**	**1**
17	110	115	97	0	0
18	95	89	110	0	0
19	85	94	85	0	0
20	95	97	94	0	0
21	95	94	103	0	0
22	105	94	97	0	0
23	110	127	107	0	0
24	95	103	97	0	0
25	110	118	97	0	0
26	105	118	118	0	0
27	85	91	94	0	0
28	115	135	110	0	0
29	**80**	94	91	**1**	0
30	105	115	100	0	0
31	90	103	91	0	0
32	100	118	103	0	0
33	**80**	97	88	**1**	0
34	90	106	97	0	0
35	105	115	110	0	0
36	90	112	100	0	0
37	90	**77**	94	**1**	**1**

6.3 Feature-based approach

In order to automate the proposed manual EEG grading (Pavlidis et al., 2019), it is necessary to mimic the process of clinical decision-making. This has been achieved by quantifying the EEG with various features that characterise different aspects of the developing brain. More specifically, in addition to a previously used set of 16 EEG features (Chapter 5), an additional set of 15 features were computed. These features include information about the temporal and

spatial organization of the EEG and have been used for the visual EEG grading (Pavlidis et al., 2019).

EEG spatial organization

The spatial organization of the EEG was estimated using the inter-hemisphere synchrony of the left and right hemispheres. EEG synchronization is representative of the brain connectivity and can be characterised by similar EEG activity occurring simultaneously in both hemispheres. It was previously reported that inter-hemispheric synchrony increases with maturation (Pavlidis et al., 2017) and therefore can be indicative of normal brain development. To estimate the level of EEG synchrony, the channels of each separate hemisphere were concatenated together. In this study, the left hemisphere of the brain was quantified with channels F4-C4, C4-O2, T4-C4, and C4-Cz, while the right hemisphere was represented with channels F3-C3, C3-O1, Cz-C3 and C3-T3 (Figure 2.5). The summary measures of EEG sub-band powers across the left and right hemispheres obtained using median and standard deviation (SD) are then contrasted. The final synchrony feature is computed as the difference between the summary measures of the left and right hemispheres.

EEG temporal organization

Preterm EEG recordings can be characterized by its discontinuity with distinctive patterns of high-amplitude burst (or spontaneous activity transients) followed by lower-amplitude periods (inter-bust intervals, IBI). The ability to maintain sleep-wake cycling is a good sign of neurological well-being. With maturation, IBI begin to shorten until reaching fully continuous EEG activity (Vecchierini et al., 2007). The temporal organization of the EEG was quantified based on the discontinuous activity using the automatic burst detection proposed by O'Toole et al., (2017). The proposed method combines multiple features of amplitude and spectral content extracted from the 1-second and 2-second long windows respectively, shifted with a 75% overlap. The feature selection is implemented using the maximum-relevance-minimum-redundancy method. The final feature set is comprised of 8 features out of the total 26 considered. The linear-kernel SVM is then used to classify each sample as either burst or inter-burst. The performance of the classifier was assessed using subject-independent LOO scheme performed on a dataset of 36 preterm neonates (GA: 23 – 30 weeks). SVM classifier outputs a sequence of ones and zeros of the same length as input data, which corresponds to the presence or absence of bursts, respectively. Prior to applying the burst detector, the EEG recordings were pre-processed by filtering and downsampling to 64 Hz. The burst detector was applied to each channel separately. An example of the EEG trace along with the corresponding output of the burst detector is represented in Figure 6.2. The mean IBI duration

for the given EEG trace is about 4 seconds, which lies within a normal range for preterms with the GA of 35 weeks (Pavlidis et al., 2017).

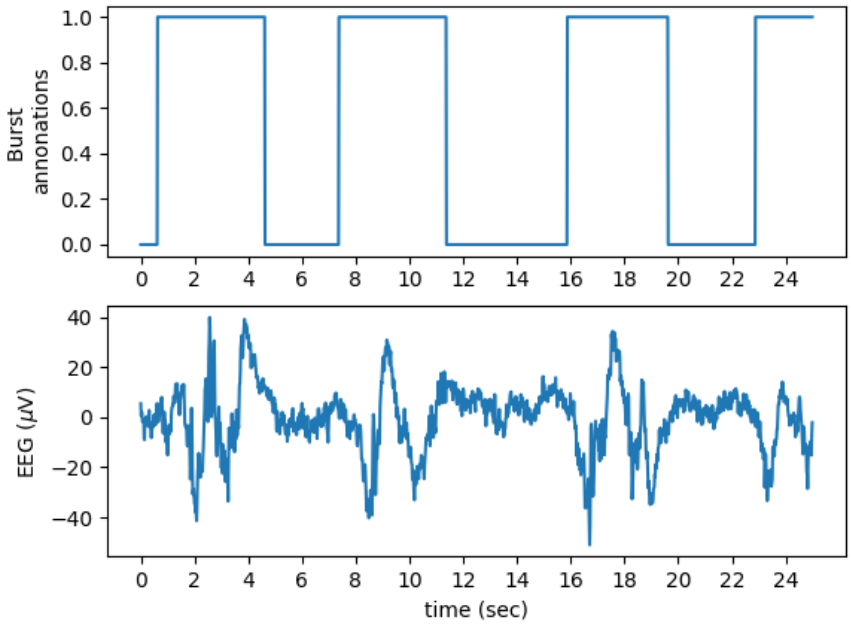


Figure 6.2: An example of the 25-second long trace of burst annotations for one channel of raw EEG (T4-C4, fs=64 Hz) from one healthy neonate. Bursts are annotated with ones, whereas IBIs are marked with zeros. The mean IBI duration for the given EEG trace is ~ 4 seconds, which is within a normal range for the preterm with the GA of 35 weeks (Pavlidis et al., 2017).

From the output of the burst detector the following summary measures are extracted: 1) the number of bursts over 60-second intervals; 2) the percentage of bursts over 60 seconds; 3) the median, the 5th and the 95th percentile of the IBI duration computed over a 5-minute window (as IBI can extend beyond 60 seconds).

Summary of extracted features

The EEG signal was bandpass filtered to the range of 0.3-30 Hz and then down-sampled to 64 Hz. The signal in each channel was segmented into 1-minute epochs with a 10-second shift. Corrupted epochs were discarded from further analysis. The preterm EEG was quantified using 31 features extracted from the time and frequency domains as shown in Table 6.2. Time domain EEG features included activity, mobility, complexity (Hjorth, 1970), the number of zero crossings along with features quantifying the discontinuity of the preterm EEG. Frequency domain features were calculated in four different EEG frequency bands: 0.3–3 Hz, 3–8 Hz, 8–15 Hz and 15–30 Hz. For each sub-band, the following features were obtained: absolute power, relative power (normalized by the total power), spectral entropy and symmetry measures (median and SD). Each EEG feature was then quantified as the median value across eight bipolar channels.

Table 6.2: Frequency- and time-domain EEG features.

Domain	Features
Time	Hjorth parameters: - activity, - mobility, - complexity, Zero crossings; Discontinuous activity: - number of bursts (burst_num), - burst percentage (burst_ratio), - median IBI duration (IBI_median), - the 5th percentile of the IBI duration (IBI_5perc), - the 95th percentile of the IBI duration (IBI_95perc).
Frequency	Measures in four EEG sub-bands: 0.3–3 Hz, 3–8 Hz, 8–15 Hz and 15–30 Hz: - absolute power (P), - relative power (RP), - symmetry measures (median as SYM_med and SD as SYM_std). Total power.
Information theory	Spectral entropy (SE) in four EEG sub-bands: 0.3–3 Hz, 3–8 Hz, 8–15 Hz and 15–30 Hz; Total SE.

6.3.1 Feature analysis: exploring feature predictive power

Prior to applying any data-driven technique, it is worth investigating the predictive capability of the features themselves. This will allow an insight into how the features are related to the problem of interest. The discriminative property of each feature has been evaluated using the AUC. This was carried out for: 1) the outcome based on 3 Bayley scales (*composite outcome*), and 2) a single scale *language outcome*.

Table 6.3 indicates that while the predictive power is quite variable across the features, several EEG features are very relevant for the task with the AUC for some reaching as high as 0.83. When comparing the predictive power of features computed on the 1-minute, 10-minute, 20-minute long epochs and the AUC obtained from the whole EEG recording, it can be seen that the feature predictive capability increases with increasing epoch length. The measures of EEG discontinuity are the features which clearly benefit from the higher amount of EEG temporal data available for the outcome prediction. The AUCs of the 5th percentile of IBI (IBI_5perc), for instance, increased from 0.63 to 0.77 for the composite outcome and from 0.61 to 0.73 for the language outcome, when comparing AUC values computed on the 10-minute epoch and on the whole EEG recording. It is known that the IBI duration decreases with maturation,

however for unhealthy preterms where the IBI duration may remain long, these could possibly have been better detected using a long EEG trace, as opposed to the short 10-minute epoch. Overall, the availability of more data showed improved generalisation across the subject.

Depending on the ground truth (language or composite outcome), some features exhibit contradictory behaviour. For instance, the number of zero crossings feature resulted in a random discrimination with AUC=0.5 for the *composite outcome*, whereas its AUC increased to 0.64 for the *language outcome* when considering 10-minute epochs. Other features like the EEG sub-band power (15-30 Hz) resulted in more consistent behaviour with an AUC of ~0.7 achieved for both outcomes, based on 10-minute epochs. Overall, according to the single feature statistics in Table 6.3, the four sub-band powers resulted in the highest AUC values.

Table 6.3: The predictive power of the EEG features with respect to 2-year composite (3 scores) and language outcomes is measured using the AUC. Results are provided for the dataset of 37 preterm neonates. The highest AUCs are in bold. The AUC was quantified for various epoch lengths (1 min, 10 min, 20 min, and the whole recording).

EEG feature		AUC (1 min) 24296 epochs		AUC (10 min) 2401 epochs		AUC (20 min) 1180 epochs		AUC (whole recording) (37 values)	
		3 scores	**Lang**	**3 scores**	**Lang**	**3 scores**	**Lang**	**3 scores**	**Lang**
1	**Power (0.3-3 Hz)**	0.65	0.61	0.66	0.62	0.68	0.64	0.77	0.71
2	**Power (3-8 Hz)**	0.68	0.63	0.69	0.63	0.69	0.64	0.77	0.7
3	**Power (8-15 Hz)**	0.68	0.66	0.69	0.67	0.7	0.67	0.73	0.71
4	**Power (15-30 Hz)**	**0.69**	**0.69**	**0.71**	**0.7**	**0.71**	**0.71**	**0.83**	**0.81**
5	**Total power**	0.68	0.64	0.69	0.65	0.7	0.66	0.8	0.75
6	**RP (0.3-3 Hz)**	0.55	0.54	0.56	0.54	0.56	0.54	0.62	0.59
7	**RP (3-8 Hz)**	0.55	0.58	0.56	0.59	0.57	0.59	0.62	0.67
8	**RP (8-15 Hz)**	0.57	0.55	0.58	0.56	0.59	0.56	0.69	0.63
9	**RP (15-30 Hz)**	0.5	0.55	0.51	0.56	0.51	0.57	0.59	0.69
10	**Total SE**	0.55	0.53	0.55	0.53	0.55	0.53	0.63	0.59
11	**SE (0.3-3 Hz)**	0.53	0.54	0.54	0.54	0.54	0.55	0.55	0.55
12	**SE (3-8 Hz)**	0.56	0.53	0.58	0.54	0.59	0.54	0.62	0.56
13	**SE (8-15 Hz)**	0.56	0.58	0.59	0.61	0.6	0.62	0.68	0.72
14	**SE (15-30 Hz)**	0.61	0.61	0.62	0.63	0.62	0.63	0.71	0.72
15	**Activity**	0.63	0.64	0.63	0.64	0.65	0.65	0.75	0.7
16	**Mobility**	0.58	0.53	0.59	0.53	0.6	0.53	0.63	0.52
17	**Complexity**	0.57	0.56	0.58	0.56	0.59	0.57	0.71	0.62
18	**Zero crossing**	0.5	0.63	0.5	0.64	0.5	0.64	0.56	0.72
19	**burst_num**	0.59	0.57	0.61	0.59	0.61	0.59	0.69	0.68
20	**burst_ratio**	0.59	0.57	0.61	0.59	0.61	0.59	0.69	0.68
21	**IBI_median**	0.62	0.6	0.62	0.6	0.62	0.6	0.75	0.71
22	**IBI_5perc**	0.62	0.6	0.63	0.61	0.64	0.61	0.77	0.73
23	**IBI_95perc**	0.61	0.6	0.62	0.6	0.62	0.6	0.71	0.69
24	**SYM_med(0.3_3Hz)**	0.57	0.55	0.59	0.54	0.61	0.54	0.59	0.51
25	**SYM_med(3_8Hz)**	0.6	0.55	0.61	0.55	0.63	0.56	0.56	0.55
26	**SYM_med(8_15Hz)**	0.57	0.55	0.58	0.56	0.59	0.56	0.6	0.61
27	**SYM_med(15_30Hz)**	0.52	0.52	0.53	0.52	0.54	0.52	0.54	0.57
28	**SYM_std(0.3_3Hz)**	0.52	0.53	0.53	0.54	0.54	0.54	0.53	0.51
29	**SYM_std(3_8Hz)**	0.55	0.52	0.55	0.51	0.55	0.5	0.61	0.55
30	**SYM_std(8_15Hz)**	0.53	0.51	0.54	0.51	0.53	0.5	0.61	0.57
31	**SYM_std(15_30Hz)**	0.53	0.54	0.54	0.55	0.54	0.56	0.61	0.63

The results of the discriminative capacity of GA and the birth weight are presented in Figure 6.3. These clinical features have shown rather low predictive power with respect to the composite outcome score (with an overlap between the classes of healthy and unhealthy groups) if compared to the physiological features computed on the whole recordings (last column, Table 6.3). In this work, the AUCs have increased from 0.55 to 0.65 and from 0.65 to 0.71 for the GA and weight respectively, when the predictive capacity is measured with respect to the language-based outcome.

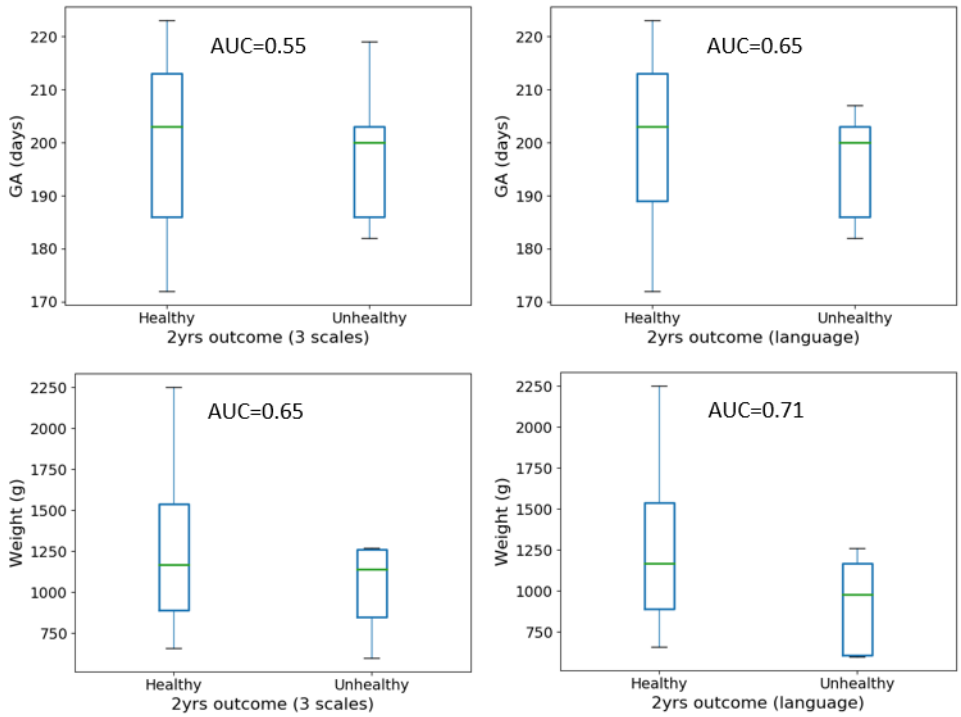


Figure 6.3: Boxplots of the clinical features with respect to 2-year outcome based on the composite of three scales and language scale.

Both the birth weight and GA are known to be survival predictors for the preterm neonates soon after birth (Draper et al., 1999). At the same time, a study by Vollmer et al., (2003) reported that the long-term outcome (assessed at 8 years of age) for preterms with an intracranial lesion depends on the type of lesion rather than GA. In the study by Temko et al., (2015) the authors developed a multimodal automated system for the neurological outcome prediction (assessed at 24 months using the Amiel-Tison method) of full-term newborns with HIE. The proposed SVM-based classifier used an initial set of 60 HR features, 57 EEG features and 3 clinical measures (Apgar score, pH and base deficit). After applying feature selection, the final feature resulted in 12 features, including only one clinical feature – Apgar score. A study by Odd et al., (2008) reported that a low Apgar score was shown to be associated with cognitive outcome. This may suggest, that clinical measures, other than GA and birth weight, may be more effective indicators of the long-term impairments – unfortunately, these were not available in our dataset.

6.3.2 Out-of-sample predictive modelling using boosted decision trees

It is worth noting that the highest predictive power of individual EEG features was achieved by averaging the feature values across each subject (resulting in one feature value per neonate) as shown in Table 6.3. The obtained predictive capability (as quantified by the AUC) of the generated EEG features can be compared with the study by Lloyd (2019), where the author has performed a similar in-sample predictive modelling by contrasting manually generated EEG grades with the outcome. The study has reported an AUC of 0.91 for the language outcome.

While the in-sample statistical predictive capability of each feature is computed from all the data, a small intra-subject data variability may influence the conclusion on whether the observed behaviour can generalise across various subjects. In-sample analysis is known to give an overly optimistic picture of the model's forecasting ability. Therefore, in order to assess the out-of-sample predictive capability of a combination of EEG features with respect to the 2-year neurodevelopmental outcome with the classical feature-based model, the boosted decision trees classifier was used (XGBoost).

The ability to provide a necessary treatment in a timely manner is crucial in a real-world clinical setting. Ideally, the intervention should be initiated as soon as a problem occurs. In this study, the classifier was designed to output the probability for each 10-minute epoch (5 minutes shift) in order to allow for the continuous real-time monitoring of the preterm neonate. This epoch length represents a trade-off between the accuracy and the amount of data used to generate predictions on the one hand, and the ability for real-time monitoring by providing instantaneous predictions on the other. More specifically, although the predictive capability of the EEG features obtained from the whole recording achieved the highest AUC (Table 6.3), this approach will not allow for the instantaneous diagnostic of the preterm's wellbeing which is essential in a day-to-day clinical setting.

In this study, the classifier was trained using the same subject-independent LOO model-selection routine described in Chapter 5. The prediction systems were scored with respect to their subject and epoch level accuracies. Results of the performance are presented in Table 6.4. The epoch-level performance corresponds to an instantaneous diagnostic and is lower than the subject-level performance which can be used for the prognostication.

The best performing classifier designed to use EEG features has resulted in an AUC of 0.83 for the subject level metrics. This result outperforms the in-sample single feature predictability (Table 6.3) with the highest AUC of 0.7 achieved by the EEG band power (15-30 Hz) for the language outcome prediction using 10-minute epochs. This, however, was not the case for the

composite outcome prediction, where the out-of-sample predictive modelling was lower than in-sample single feature predictive capability, AUC=0.63 vs AUC=0.71 (Table 6.3, 10-minute epochs).

Table 6.4: AUC values for the long-term outcome prediction. The outcome is represented as a composite of 3 Bayley scales and single language scale. AUCs for the best performing EEG-based system are in bold.

	Composite of 3 scales		**Language scale**	
	Epoch AUC	Subject AUC	Epoch AUC	Subject AUC
EEG features	0.57	0.63	**0.67**	**0.83**
EEG & clinical features	0.63	0.63	0.67	0.73

6.4 End-to-end deep learning for outcome prediction

6.4.1 Related work

CNN is a type of deep NN which was originally developed for images analysis as a special case of the traditional feedforward network. CNN learn relevant features using an end-to-end hierarchical representation of the data. CNNs are well-known for their great success in solving computer vision problems (Hinton et al., 2012; Krizhevsky et al., 2012). This was followed by the application of CNN to audio, music and speech processing tasks (Abdel-Hamid et al., 2014; Lee et al., 2009; Schlüter and Böck, 2014), where the time series were represented as 'images' using time-frequency decomposition. In the field of physiological signal processing, EEG spectrograms have been used for the task of epilepsy prediction in adults (Korshunova et al., 2018). Likewise, the time-frequency images obtained using the short-time Fourier transform were used together with CNN for brain-computer interfaces (Stober et al., 2015; Zhang et al., 2017).

At the same time, other studies have used raw waveforms (Hoshen et al., 2015; Tüske et al., 2014) as the CNN input. In (Lee et al., 2017) it was demonstrated that a combination of deep architectures (more than 10 layers) and small filters (2 – 3 sample long raw audio waveforms) is effective for the task of music auto-tagging. In the field of neonatal care, raw EEG data together with CNN have shown state-of-the-art results for the task of neonatal seizure detection (O'Shea et al., 2017).

A recent study (Ansari et al., 2018) has proposed an 18-layer fully-connected CNN to detect quiet sleep in healthy preterm infants (according to the Bayley Scales of Infant Development-II, mental and motor score >85) using multichannel EEG signals recorded between 27 and 42 weeks PMA. The study reported performance as measured by the area under the mean and median ROC curves of 92% and 98%, respectively. Another study by Kawahara et al., (2017) proposed a CNN called BrainNetCNN for the prediction of the Bayley-III cognitive and motor scores of development assessed at 18 months of age using the DTI of the preterm neonates.

6.4.2 Methodology

The main advantage of the CNN lies in its ability to perform end-to-end learning. This allows for replacing the feature extraction, feature selection and classification procedures with a single network. Learning from the raw data allows for a number of advantages. More specifically, the EEG is known to be a non-stationary signal while some features require a stationarity assumption by extracting features from a short time interval. At the same time, Fourier transform which allows for converting signals from the time to frequency domain assumes that the time signal is a sum of the set of sine waves of different frequencies. However, by learning from the raw signals, there is no need to make any assumptions about the data.

A study by Wulsin et al., (2011) has demonstrated that the deep belief net classifier trained using raw EEG data has outperformed the network trained on the relevant EEG features. Based on this finding and the results of the previous successful applications of CNN to raw EEG data (Ansari et al., 2018; O'Shea et al., 2017), this work investigates the ability to predict the 2-year outcome using raw EEG waveforms.

From 2D to 1D

For the problem of image processing, CNN usually incorporates 2D filters which move across the image. In this work, we are dealing with temporal multichannel EEG signals. It is known that EEG spatial information is important as it can be indicative of normal brain functioning (Pavlidis et al., 2017). While in our study the multichannel EEG is comprised of 8 channels, the number and order of channels may vary ('double banana' montage (Stevenson et al., 2019)). Therefore, in order to make the network invariant to both the number and the order of the EEG channels, the 8-channel input to the network was processed channel by channel using 1D convolutional filters. As a result, up to the final layer of CNN, the EEG is represented with all (eight) channels. Finally, the classification decision is made on the 8 parallel paths using global average pooling (GAP) followed by a softmax function.

EEG pre-processing and segmentation

Raw EEG signal was bandpass filtered with cut-off frequencies of 0.3 and 30 Hz. This signal was then downsampled to 64 Hz. The EEG is then split into 2-minute epoch shifted every second. This has resulted in the input dimension of $7680 \times 8 \times 1$. Here $W = 7680$ is epoch length, $H = 8$ corresponds to eight bipolar EEG channels, and depth $D = 1$.

Small filters

Following the studies which have successfully applied the CNN to temporal signals (Lee et al., 2017; O'Shea et al., 2017), in this work we have also used small, 4-sample long convolutional filters. Beginning with a small information gathering filter size allows for learning an increased range of filter depths and thus, capturing smaller and more complex features. Each convolutional layer represents a higher level of feature abstraction of the previous layer. Smaller filters are also more computationally efficient as they reduce the number of weights (parameters) that the network needs to learn.

Fully convolutional network

Fully convolutional NN corresponds to the network which does not rely on any fully connected (FC) layers. In an FC (dense) layer every input is connected to every output, and each connection has its own weight. This is followed by the nonlinear activation function. FC layer results in $(n_{inp} + 1) \times n_{out}$ weights, which is usually a computationally expensive step due to the possibly high number of parameters. The FC network architecture has been widely used for various classification tasks (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014). Typically, the feature selection stage is realised using a number of convolutional and pooling layers. This is then usually followed by the classification step, where FC layers learn to classify every input of the network based on the output of the last convolutional layer flattened into a 1D array. VGG-net, the winner of the ImageNet Challenge used FC layers for the classification step (Simonyan and Zisserman, 2014). An example of a typical CNN/FC network architecture is represented in Figure 6.4.

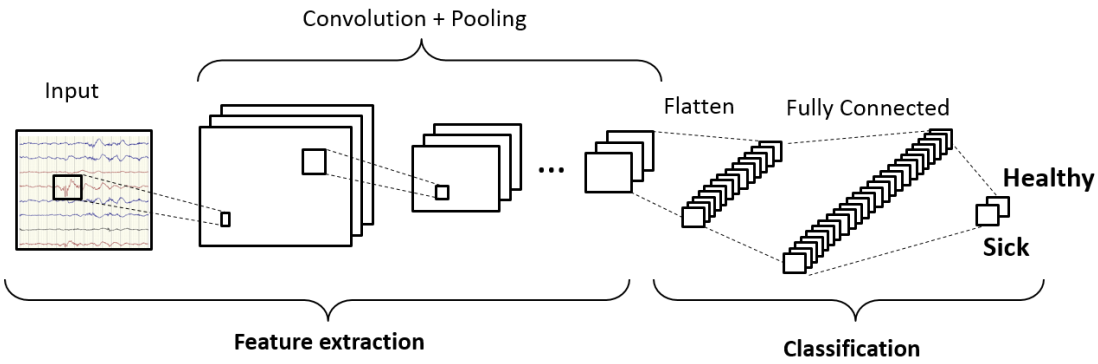


Figure 6.4: An example of CNN with fully connected layers.

A fully convolutional network is composed of convolutional layers only, without any FC layers. In contrast to the FC layer, the neurons of the convolutional layer are connected only to nearby neurons in the previous layer (Figure 6.5). This provides a dramatic reduction in the number of trainable parameters. Fully convolutional NNs, therefore, are less prone to overfitting and require fewer regularization routines, such as dropout. The features maps in the final layer of a fully convolutional network are directly connected to the final class values

(Figure 6.5). The fully convolutional network makes decisions based on the final weight matrices by looking for a particular class in the input. These networks allow the prediction of output from an arbitrary-sized input (Shelhamer et al., 2017). All weights of the network are spatially invariant convolutional weights, therefore, there are no requirements on the specific shape of the input.

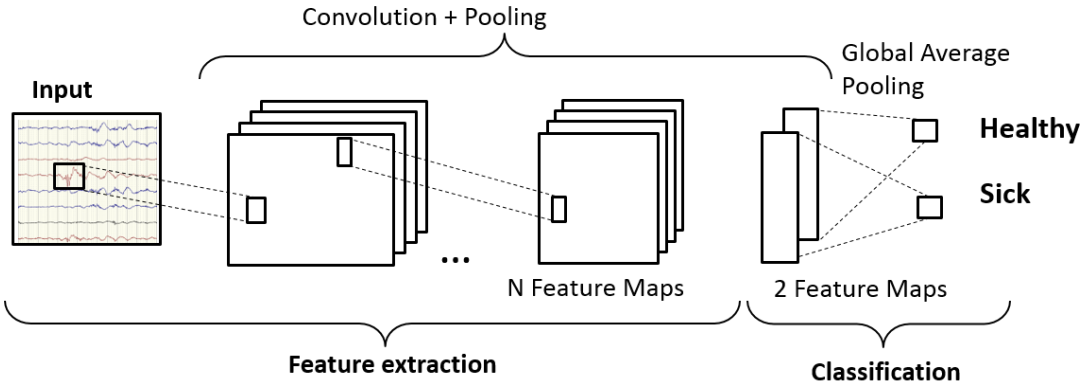


Figure 6.5: An example of fully convolutional network architecture.

Network architecture for outcome prediction

In this work, the network architecture was designed to be fully convolutional, without any fully connected layers as shown in Figure 6.6. The network was constructed using the open source Keras deep learning library (Chollet, 2015). Due to the high input dimensionality and small filter size, the pooling layers were used in order to improve the receptive field and reduce the number of parameters. The proposed architecture incorporates residual blocks (Figure 6.7), which allow for training deeper network by using skip connections.

Two feature maps in the final convolutional layer act as a classifier. The average value (GAP) taken across each final feature map produces 2 class values (x_i, for i=2), which are converted into a set of probabilities for healthy and poor outcomes using softmax operation as follows:

$$softmax(x_i) = \frac{e^{x_i}}{\sum_i e^{x_i}} \tag{6.1}$$

Although it is a common practice to use FC layers in order to perform classification on the extracted features (Ansari et al., 2018; He et al., 2015a; Simonyan and Zisserman, 2014), the GAP of the two final feature maps learns to automatically correspond to healthy and unhealthy classes. The receptive field of the first layer is 4 samples, whereas in the last layer the receptive field is increased to 6520 samples, which corresponds to 102 seconds of raw EEG (sampled at 64 Hz) (Dumoulin and Visin, 2016). This suggests that high frequency components (high pass filter) are learnt in the first layers where the receptive fields are narrow. Likewise, the deeper

layers with wider receptive fields learn the low frequency components (low pass filter) of the EEG signal.

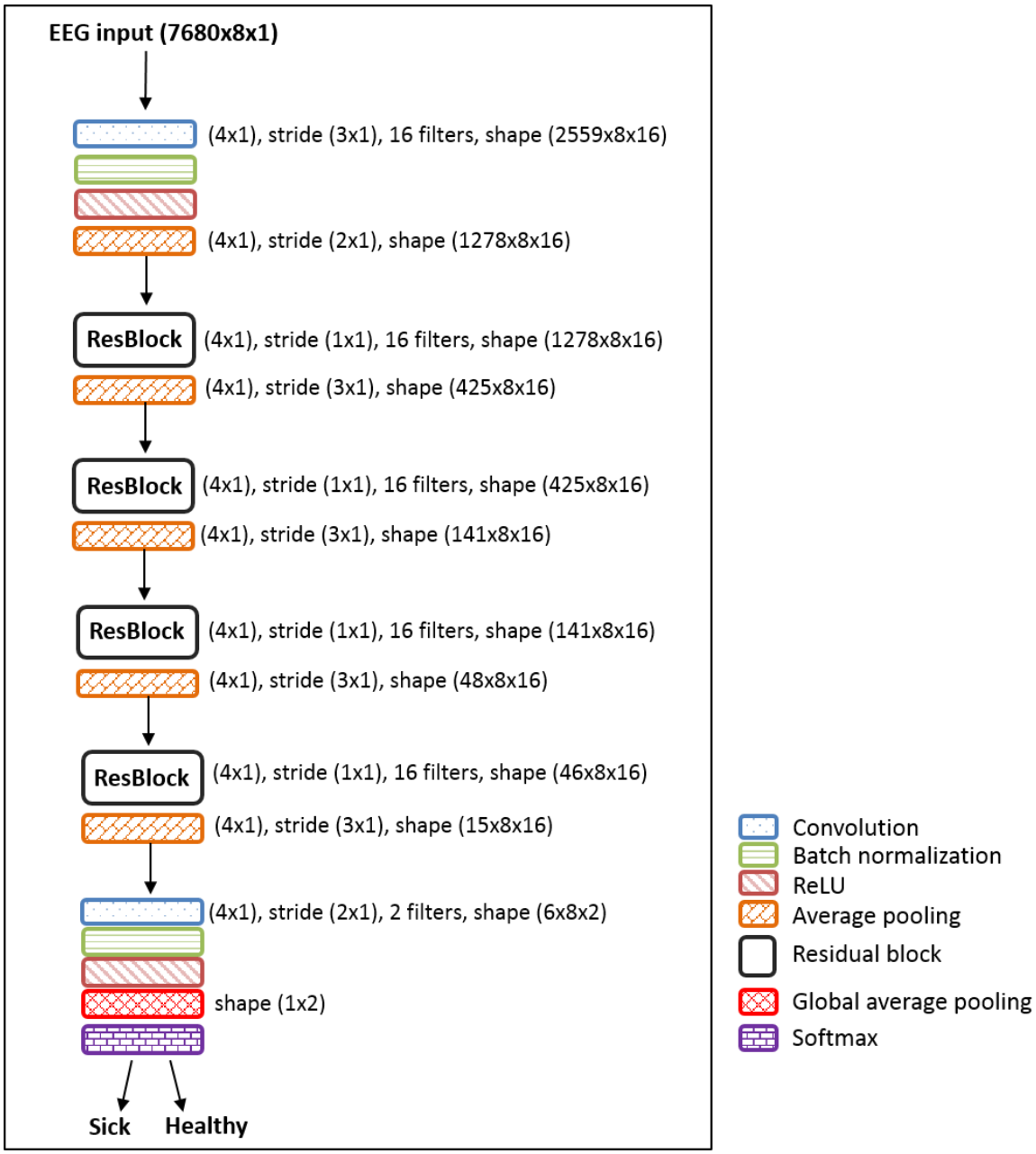


Figure 6.6: Schematic diagram of the fully convolutional neural network (16 filters) used for the 2-year language outcome prediction.

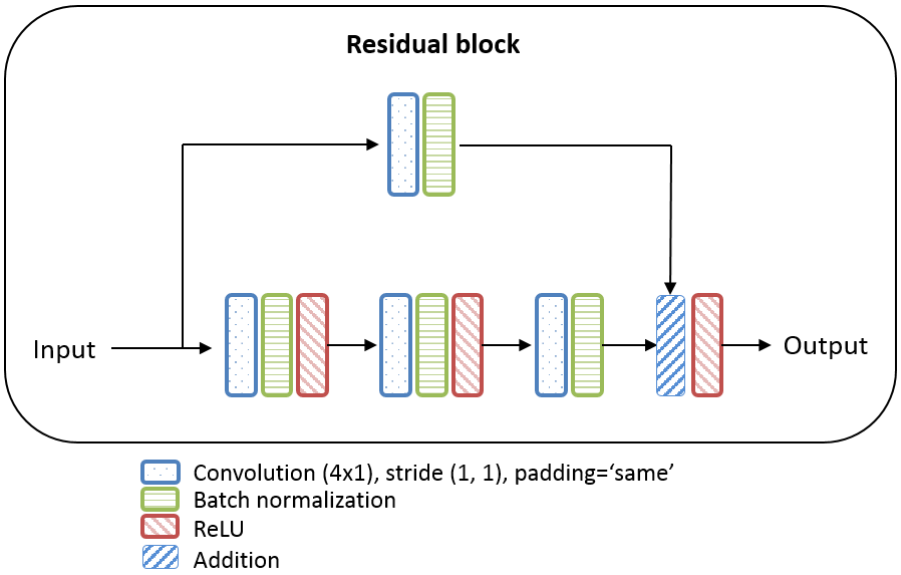


Figure 6.7: Schematic representation of the residual block within the network from Figure 6.6.

6.4.3 Optimization and classification performance

The network was trained using stochastic gradient descent with a learning rate of 0.001 and a batch size of 1024. Categorical cross-entropy was used as the loss function. For regularisation, batch normalization was applied after each convolutional layer.

The AUC was similarly used to measure the performance of the CNN. The LOO subject independent performance assessment is used in this work to estimate the generalisation error. Unlike the feature-based classifier, where parameters are selected during an internal CV routine by splitting training data into the train and internal CV test sets, the CNN network was trained on all data available, N-1 subjects, where N is a total number of neonates in the dataset. This is done in order to maximise the number of training examples and provide the network with diverse data. Depending on the number of filters in the network, the number of training epochs was set to 50 (for 8 filters) and 30 (for 16 filters). No early stopping criteria were found to be necessary for this study. In order to address the problem of unbalanced classes (30 vs 7 for the language outcome), minority class data repetition was applied resulting in a total of 326K training examples. As a result, during training, the classes in the mini batches were balanced. The classification performance for the *language outcome* prediction resulted in an epoch level AUC of 0.62 and subject level AUC of 0.81 using the network with 16 filters. At the same time, the CNN with 8 filters resulted in AUC=0.61 and AUC=0.78 for the subject- and epoch-level performances correspondingly. An example of the training learning curve (network trained on all the data) is represented in Figure 6.8. It can be seen that while the model gets a reasonable fit of the training data after 30 iterations, a small variation in the performance can be observed for different LOO iterations.

Table 6.5 represents the AUC values obtained for different initialization points. The final AUC is then computed by averaging test predictions of models with different parameter initialisation.

Table 6.5: AUC for the long-term language outcome prediction using the proposed CNN architecture. The AUCs obtained using the averaged predictions generated by the network with different initialisations are in bold.

	Language scale (8 filters) 4882 parameters		**Language scale** (16 filters) 17398 parameters	
	Epoch AUC	Subject AUC	Epoch AUC	Subject AUC
Initialisation 1	0.58	0.65	0.59	0.74
Initialisation 2	0.61	0.79	0.6	0.75
Averaged probabilities	**0.61**	**0.78**	**0.62**	**0.81**

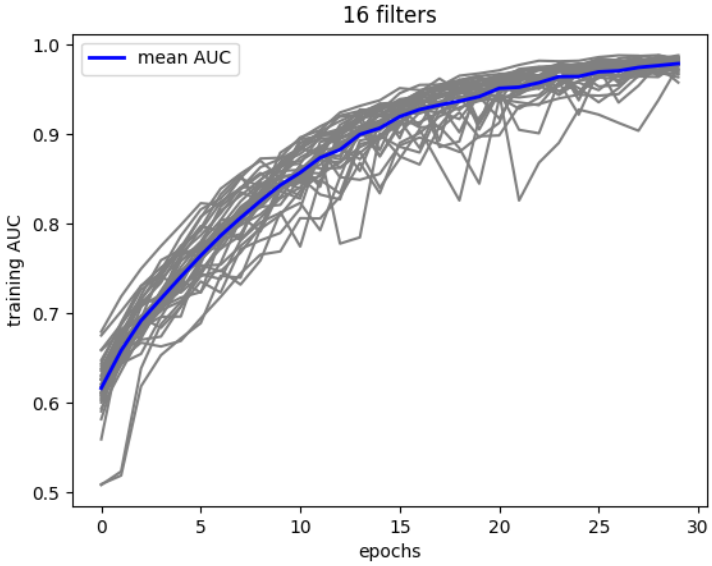


Figure 6.8: An example of the training AUC for each LOO iteration (grey), and its corresponding mean AUC (blue) for the network with 16 filters.

6.5 Discussion

The prediction of the long-term outcome based on the early EEG is a challenging task due to the complexity of the developing brain (Pavlidis et al., 2017) as well as a number of possible confounding factors which may potentially influence the developmental process (Brown et al., 2014). Results obtained in this study indicate that for a given dataset the classical feature-based classifier has outperformed the end-to-end learning approach with subject-level AUC of 0.83 vs 0.81. To the best of our knowledge, this is a first application of the deep learning applied to the raw EEG data for the problem of 2-year (language) outcome prediction. The CNN architecture used in this study was designed based on the results of other studies of audio

analysis with sample sized convolutions (Lee et al., 2017), and fully convolutional networks applied to term neonatal EEG signals for seizure detection (O'Shea et al., 2017).

The CNNs have been previously applied to the raw EEG for the task of sleep detection in preterms (Ansari et al., 2018) and seizure detection in term neonates (O'Shea et al., 2017). Both sleep and seizure detection implies variability of the data within the patient, where the same subject can have episodes of different events (e.g. seizure and non-seizure, quiet sleep and active sleep). These problems, therefore, can be defined as event detection tasks. In our study, however, the whole recording of each neonate belongs to either one of two classes with no intermediate events defined. In other words, the lack of inter-patient data variability in a population of 37 newborns makes it more difficult to generalise across subjects.

In order to maximize the variability and the amount of the training data, the commonly used validation set for the purpose of model selection was not incorporated for the end-to-end learning approach. Instead, the model selection routine was carried out separately from the LOO performance assessment. The choice of network architecture was performed by training on 35 neonates, while excluding one sick and one healthy neonate for evaluation purposes. In this way, other network architectures including larger filter sizes, fewer convolutional layers (shallow network), and varying length of the input EEG epoch were tested. The best network architecture was selected empirically based on the predictions of two (one sick and one healthy) left out subjects. The absence of a reasonably sized validation set is a potential limitation of the study which should be addressed when a larger labelled dataset of preterm EEG becomes available.

The lack of data variability and absence of any domain-based information has made the end-to-end learning approach less suitable for this particular task. The classical feature-based approach, on the other hand, has benefited from the domain-based information provided in the form of the choice of EEG features which quantify important aspects of the premature brain. This allowed for focusing on the relevant information for better generalisation across the subjects.

Noisy labels

It worth mentioning that the majority of poor 2-year outcomes were contributed by the language scale (Table 6.1). The Bayley scales (Bayley, 2006) is a test initially designed to assess the developmental functioning, including language, for infants and toddlers whose first language is English. Lowe et al., (2013) reported that preterm neonates whose primary language was Spanish had similar cognitive, but lower language scores as compared to those whose primary language was English. The authors concluded that the native language spoken

at home should be taken into consideration. An association between the bilingual environment and lower mental development of the Bayley Scales in VLBW infants was reported by Walch et al., (2009). Another study (Spencer-Smith et al., 2015) conducted on 105 children born <30 weeks GA concluded that cognitive and language scales at 2 years were not strongly predictive of impairments at 4 years of age. Therefore, it is possible to conclude that our labels are prone to the subjectivity of observers and several other confounding factors, including native language spoken at home. As a result, one might argue that the labels used in this study are weak labels which may contain noise (error). A study by Guan et al., (2017) demonstrated an interesting result by training on the MNIST dataset with inaccurate labels. The authors reported a test error rate of 1.01% when using all correct labels and only 2.29% error rate when a randomly selected 50% of labels were corrupted. It has been concluded that in order to get a comparable performance using noisy labels as compared to clean labels, more training examples for noisy labels are required. Therefore, one of the ways to address the problem of possibly noisy labels used in our study is by getting more training examples (i.e. more subjects). This once again highlights a need for a larger labelled dataset.

A study by Kawahara et al., (2017) proposed a CNN called BrainNetCNN for the prediction of the Bayley-III cognitive and motor scores of development assessed at 18 months of age by using the DTI of the preterm neonates as an input. The DTI is an MRI-based neuroimaging technique which allows for estimation of the location and orientation of the brain's white matter. The study was conducted on the population of preterm neonates; the DTI images were acquired between 27 and 46 weeks PMA. The authors reported that motor scores were predicted more accurately than cognitive scores, with a mean absolute error between the actual and predicted Bayley-III motor composite score of 11%. While DFI allows for analysis of the microstructure of the white matter in 3D space, the EEG is used to obtain information about the brain activity with a temporal resolution. The combination of both, spatial and temporal techniques, may, therefore, allow for much deeper insight into the brain functioning and connectivity.

In the study (Lloyd, 2019), the author represented 2 hours of EEG recording with a single value of manually generated EEG grade (Pavlidis et al., 2019) while performing in-sample predictive modelling. In our work, however, we have performed an out-of-sample predictive modelling using ML techniques.

Imbalanced classes

The problem of class imbalance in the training data is very common, especially in the medical field (Mac Namee et al., 2002). This is due to the fact that usually sick or at-risk patients are monitored, whereas any intervention with a healthy neonate is avoided, if possible. As a result,

labelled data for classification is usually restricted to favour a certain class. This leads toward the prediction of the more frequently occurring outcome in the training data. In this study, the class imbalance was addressed by repeating data of minority class.

The class imbalance should be taken into account when selecting the model assessment technique. While the accuracy measure reflects the underlying class distribution, the area under the ROC curve used in this study, is robust to class imbalance, as it plots true positives against false positives at various thresholds.

Administered medication

Administered medication can suppress physiological activity. Preterm EEG, for instance, may exhibit increased discontinuous activity with prolonged IBI (Bell et al., 1993). Pavlidis et al., (2019) accounted for a possible effect of medications when manually grading preterm EEG. In this thesis, however, no information was available regarding the administered drugs, which is another potential limitation of the study.

6.6 Conclusion

The results obtained in this study demonstrate the feasibility of the construction of a computer-based decision support system using ML techniques for the prediction of long-term neurodevelopmental outcome. More specifically, for the available dataset, a classical feature-based approach for the problem of the long-term neurodevelopmental outcome prediction has outperformed an end-to-end deep learning technique with the subject-level AUCs of 0.83 vs 0.81, correspondingly. It is essential to acknowledge the limitation of the study, including the lack of variability in the data and possible noise within labels. Therefore, further confirmatory research needs to be conducted in order to validate the obtained results on a larger dataset of preterm EEG recordings. An increased presence of labels with poorer scores (<85) for motor and cognitive scales (not language only) would also help to support the finding of this work and to provide an insight into the association of early brain function and different aspects of further development. Overall, the obtained findings present a first and promising step toward the development of an automated system for the early (before the discharge from the NICU) EEG-based assessment of long-term neurodevelopmental well-being for premature neonates.

Chapter 7: Conclusions and future work

Preterm neonates are at high risk of various complications which lead to the death of more than one million children annually. Modern NICUs allow for a continuous flow of information in the form of physiological signals, which can be used to assess the health status of the neonate and to prognosticate possible complications. Clinical decisions based on such information overload are prone to mistakes due to a possible lack of domain-specific knowledge, especially when the problems are difficult, if not impossible, to identify with the naked eye; fatigue-related mistakes are also frequent in the real-world clinical settings, when the high volume of information must be analysed under a tight time constraint.

Computer-based analysis of physiological signals allows for faster, accurate and objective decision making at the cot-side. In this context, the main objective of this thesis is to use the available data in the NICU to provide more efficient management of preterm neonates. This has been achieved by developing an automated system which can assist clinicians with diagnostic and treatment-related decisions for preterm neonates.

This chapter will describe the contributions of the thesis, in relation to the main objective, along with subsequent conclusions and limitations. Finally, the future directions of the research in this area are proposed.

7.1 Concluding summary and contributions

The research conducted in this thesis focused on the development of a decision support system for the intelligent monitoring of preterms in the NICU. In this work, the well-being of the neonate has been quantified at different stages of life. An early assessment conducted in the first 12 hours of life (CRIB score) was followed by the short-term outcome obtained at discharge from the NICU. At 2 years of age, the Bayley scales were used to quantify the long-term neurodevelopmental outcome. Being able to properly address the health complications at any stage of life is essential for the efficient management of the vulnerable population of preterms, and for their future well-being. This idea has served as the primary motivation for the research conducted in this thesis. More specifically, one by one, in a chronological manner, the thesis has addressed the neonatal health complications which might arise at different stages of an infant's life. This section will provide a brief summary and contributions of each chapter.

Chapter 2 provided an overview of the medical background of the health and complications that usually occur soon after birth in preterm neonates. The main physiological signals which are usually available in the NICU, such as EEG, ECG, and BP are described. This information gives a better understanding of the current treatment techniques and an appreciation of the challenges clinicians face when dealing with the vulnerable preterm neonatal population. The details of physiological and neurodevelopmental health measures used in the thesis are also provided.

An overview of the technical methods, including signal processing, feature extraction and measures of coupling between physiological signals used in this thesis, were presented in **Chapter 3**. It was shown that unlike classical methods of linear coupling, such as correlation and coherence, the nonlinear information measures based on entropy may be more suitable when dealing with complex physiological signals such as EEG. This chapter also introduced the main ML concepts along with a more detailed review of the supervised ML methods, which are most frequently used in the medical field (Deo, 2015). This information enabled the selection of a suitable set of techniques to address each neonatal health complication investigated in this thesis.

Chapter 4 investigated the coupling between physiological signals recorded in the NICU and explored how this interaction was affected by the early health status of the preterm as assessed by the CRIB score. The chapter provided a detailed explanation of the interaction modelling between the cortical activity, quantified by EEG features, and mean arterial blood pressure (MAP). The level of coupling between two physiological systems was assessed in the context of the health status of the preterm. In order to ensure the reliability of the obtained results, an appropriate null hypothesis was tested for every computed measure of interaction.

Obtained findings demonstrated that the physiological reaction to the changes in BP is associated with lower risks to the preterm, which is the main contribution of this study. More specifically, the results have shown that a higher risk of neonatal mortality (higher CRIB scores) is associated with a lower level of nonlinear interaction between EEG and MAP. Our findings suggest that nonlinear measures of interaction are more suitable when measuring coupling between brain function and blood pressure.

While several studies have previously investigated the causal relationship between brain functioning (EEG) and changes in cerebral oxygenation (NIRS) (Caicedo et al., 2016; Roche-Labarbe et al., 2007), to the best of our knowledge this is the first study to investigate the interaction and the directional information flow between MAP and EEG for preterm neonates. It was demonstrated that higher CRIB scores are associated with higher levels of information flow from EEG-to-MAP as measured by TE. This allows us to hypothesise that the normal

wellbeing of a preterm neonate can be characterised by a strong nonlinear coupling between brain activity and MAP, whereas the presence of weak coupling with distinctive directionality of information flow may be associated with an increased risk of illness severity in preterms.

There is still no clear definition of hypotension in preterm neonates and the decision on whether it should be treated remains disputed with considerable variability across clinical practice (Dempsey and Barrington, 2006). The findings obtained in Chapter 4 can potentially contribute towards the generation of a hypothesis in the field of hypotension management in preterms and the interrelation between cerebral activity and BP. More specifically, two channels of EEG are now routinely used in infants who are suspected of having a brain injury. Therefore, we anticipate that a module in a bedside monitor incorporating the algorithm to compute and visualise the measure of interaction between BP and cerebral activity would be feasible. It could allow for real-time decision support for more efficient management of hypotension in preterm neonates. Therefore, instead of relying solely on generic MAP<GA threshold for the definition of hypotension, the level of interaction between two systems could help to guide clinicians to provide patient-specific treatment.

Following the chronology of the assessment of preterm's well-being, **Chapter 5** concentrated on an investigation into the relationship between different physiological modalities (EEG, ECG, and BP) and the short-term health outcome of the preterm quantified by the CCS at discharge from the NICU. Novel EEG and HRV based decision support tools for the continuous estimation of the probability of neonatal morbidity using the observed multimodal physiological data and a boosted decision trees classifier have been proposed.

First, the predictive capability of HRV and EEG for estimating the short-term health outcome was assessed and the predictive power of both HRV and EEG features during episodes of low BP was studied. The obtained findings indicated that the predictive power of extracted features increases when observed during low BP episodes; with a single best HRV feature leading to an AUC of 0.87. More specifically, HRV of healthy preterms demonstrated a significant reaction to the drop in BP. This result supports the finding of Chapter 4, where the increased level of interaction between EEG and BP in preterms was shown to be associated with lower risks of illness severity.

On the next stage, an ML approach was utilised to perform the out-of-sample predictive modelling of EEG and HRV features. It was shown that a predictor of short-term health based on EEG features performed worse than one based on HRV features. Combining multiple HRV features with a boosted decision trees classifier resulted in an AUC of 0.97, using a LOO patient independent performance assessment. The developed decision support system allows

for the continuous estimation of the probability of neonatal morbidity based on the observed physiological data.

Non-invasive techniques for BP measurement are known to be unreliable in small and sick infants. At the same time, an excessive intervention in sick preterms is undesirable and gold standard invasive BP recordings at times may be difficult to obtain (Weindling, 1989). The non-invasive ECG, on the other hand, is routinely recorded in preterms. In this context, the study identifies 3 possible decision support tools for the clinical prediction of short-term outcome:

1) Event-driven outcome predictor based on HRV and BP for use during episodes of hypotension;

2) Continuous assessment of the outcome based on HRV and BP;

3) Continuous outcome prediction based on HRV alone, in cases when invasive BP recording is not available or deemed unsafe to collect.

Each system outputs a probability of morbidity for every five-minute window providing a real-time insight into the neonate's health status. Such systems can potentially assist clinicians when monitoring preterm neonates, who may have low BP. Results of Chapter 5 present a promising step towards the use of physiological data in building an objective decision support tool for the clinical prediction of short-term outcome.

While the previous chapters have incorporated clinical scores assessed soon after birth, **Chapter 6**, focuses on the long-term outcome assessed at 2 years of age. Here an automated system for the prediction of the 2-year neurodevelopmental outcome based on the EEG recorded before the discharge from the NICU was developed. The results indicated that for an available annotated dataset, the classical feature-based approach outperformed a CNN trained using end-to-end optimization; this highlighted the need for a large dataset when using DL techniques. To the best of our knowledge, this is the first utilization of the DL applied to raw EEG data for the problem of 2-year outcome prediction. Obtained findings demonstrate the feasibility of the construction of a computer-based system using ML techniques for the prediction of long-term neurodevelopmental outcome. Early interventions for preterm neonates have previously shown a positive influence on their motor and cognitive development (Spittle et al., 2015). Results of this study can potentially contribute to the development of a system which will allow faster identification of neonates who are at risk of long-term impairments. This may ultimately help clinicians to provide the necessary treatment in a timely manner, thus allowing for an improved quality of care with potential far-reaching effect on the health service.

7.1.1 General limitations

Administered medication can suppress physiological activity as quantified by BP, ECG, and EEG. Preterm EEG, for instance, may result in increased discontinuous activity with prolonged IBI (Bell et al., 1993). Pavlidis et al., (2019) has accounted for the possible effect of medications when manually grading preterm EEG. In this thesis, however, no information was available regarding the administered drugs, which is a potential limitation of the study.

The population of preterm neonates is extremely vulnerable, and it is usually difficult to obtain permission for any kind of intervention from neonatologists. This, therefore, may introduce a bias in the available clinical data, as most of the time only babies identified as sick are monitored. In order to better represent the general neonatal population, a larger dataset should be collected. This will allow to test the generalisation of the developed systems on a larger cohort of preterm neonates.

7.2 Future work

There are several areas which can further advance the proposed automated systems for the intelligent monitoring of preterm infants in the NICU. This section will consider the possible avenues for improvements of the proposed systems.

Preterm cortical activity can be characterised by a number of maturation features (Pavlidis et al., 2017), such as continuity, sleep states, and others. In Chapter 4, we assessed the first days of life of the preterm only, where every infant was represented with a single summary measure (median across the recording). As a result, this did not allow us to investigate the possible impact of the cyclical activity, such as sleep states, on the coupling across time. Further research can potentially incorporate additional characteristics to account for the discontinuous pattern of preterm EEG while considering the temporal relationship between brain activity and BP.

Feature-based classifier

For the problem of short-term outcome prediction, as quantified with CCS, both EEG and ECG segments could be significantly corrupted by artefacts. The combination of HRV and EEG features, where both recordings were simultaneously free of artefacts, was only possible on a drastically reduced amount of data. Both EEG and ECG are extensively used for assessing aspects of newborn health. Promising results have been previously obtained for the automated computer-based outcome prediction in full-term neonates using a combination of multimodal features including HRV, EEG and clinical features (Temko et al., 2015). Therefore, a combination of preterm EEG and HRV modalities in a single classifier could be researched in

the future in order to further improve the performance of outcome prediction and make it more robust to artefacts.

The HRV characteristics used in this thesis are the features which were identified in the literature for full-term and preterm HR analysis (Selig et al., 2011; Temko et al., 2015). These features, however, have not been previously studied with respect to BP. Therefore, future research in the area of preterm HR analysis can potentially improve the representation of the preterm heart activity by expanding currently utilised feature set with more sophisticated /suitable HR characteristics.

End-to-end learning

Depending on the type of ML technique, various approaches for further improvements can be proposed. While for the classical feature-based system used in Chapter 5, further research in the field of feature engineering might be beneficial, the end-to-end DL models would require a different approach. A major challenge in the development of a robust model based on physiological signals lies in the availability of a sufficient amount of informative labelled data obtained for each patient. This aspect is especially crucial when training deep NN classifiers. The lack of data variability, the high dimensionality of the input EEG, as well as the choice of network architecture for EEG classification were the main challenges of the work conducted in Chapter 6.

Transfer learning is one of the possible techniques which can potentially help to improve the performance of the classifier along with the acceleration of the training process itself. This is achieved by reusing the weights in one or more layers from a model pretrained on a similar task (Ng et al., 2015). More specifically, for the problem of long-term outcome prediction using raw EEG from the preterm neonate, it is possible to utilise weights from a network which was similarly trained on the raw EEG. The CNN is capable of identifying patterns in the provided input, starting from simple lines and edges in the first layers, up to more complex shapes and objects when reaching deeper layers. In this context, the transfer learning technique provides the network with some low level pre-learned features and allows for an easier starting point in training.

If using convolutional network without pooling, the size of receptive field will grow linearly. This can be limiting, especially in situations where the input dimension is high. Dilating convolution techniques allow the size of the receptive fields to exponentially increase. This enables a multiscale representation of the data by combining local and global information (Yu and Koltun, 2015). The WaveNet paradigm has demonstrated the effectiveness of dilated convolutions when applied to time series (Oord et al., 2016). When compared to the current

best text-to-speech systems, the study has reported state-of-the-art performance achieved when inreasing the size of the receptive field. In the study which assessed the association of EEG with long-term outcome (Lloyd, 2019), long multichannel EEG was graded with a single grade per subject. Therefore, dilation technique can be potentially useful for learning from a longer EEG epoch. This may benefit from low frequency oscillations, which are known to be present the preterm EEG (Pavlidis et al., 2017).

The process of NN tuning is more empirical rather than theoretical. There is no 'good' architecture that solves different problems in the same optimal way. Although there are techniques which have shown to perform better, including a small filter size (Simonyan and Zisserman, 2014) and residual connections (Bai et al., 2018), the adjustment of the network should be performed gradually, while analysing the result of every change. Normally, the network capacity (e.g. number of hidden layers and filters) is increased with the intention to fit the training data, while at the same time making sure it does not overfit to the unseen validation set. If overfitting happens, different regularization techniques (e.g. L1 & L2 regularization, dropout, early stopping, data augmentation) can be used to tone it down. Network tuning continues until reaching an optimal bias-variance trade-off.

Sepsis is a very common complication in neonates occurring soon after birth, which increases the vulnerability of the brain due to white matter damage (Mallard and Wang, 2012). Alshaikh et al., (2014) reported an association between sepsis in preterm infants and increased long-term cognitive delay. The study by Kawahara et al., (2017) proposed a CNN classifier to predict the Bayley-III cognitive and motor scores in preterms by using the DTI as network input. The DTI is an MRI-based neuroimaging technique which is used to estimate the location and orientation of the brain's white matter. The combination of EEG temporal information with spatial neuroimaging can potentially provide a better insight into the brain functioning and connectivity, and their association with both short-term and long-term outcomes.

REFERENCES

Abdel-Hamid, O., Mohamed, A., Jiang, H., Deng, L., Penn, G., Yu, D., 2014. Convolutional Neural Networks for Speech Recognition. IEEEACM Trans. Audio Speech Lang. Process. 22, 1533–1545. https://doi.org/10.1109/TASLP.2014.2339736

Aboalayon, K.A.I., Almuhammadi, W.S., Faezipour, M., 2015. A comparison of different machine learning algorithms using single channel EEG signal for classifying human sleep stages, in: 2015 Long Island Systems, Applications and Technology. Presented at the 2015 Long Island Systems, Applications and Technology, 1–6. https://doi.org/10.1109/LISAT.2015.7160185

Ahmed, R., Temko, A., Marnane, W., Lightbody, G., Boylan, G., 2016. Grading hypoxic–ischemic encephalopathy severity in neonatal EEG using GMM supervectors and the support vector machine. Clin. Neurophysiol. 127, 297–309. https://doi.org/10.1016/j.clinph.2015.05.024

Akselrod, S., Gordon, D., Ubel, F.A., Shannon, D.C., Berger, A.C., Cohen, R.J., 1981. Power spectrum analysis of heart rate fluctuation: a quantitative probe of beat-to-beat cardiovascular control. Science 213, 220–222.

Aletti, F., Hammond, R., Sala-Mercado, J., Chen, X., O'Leary, D., Baselli, G., Mukkamala, R., 2013. Cardiac Output is Not a Significant Source of Low Frequency Mean Arterial Pressure Variability. Physiol. Meas. 34. https://doi.org/10.1088/0967-3334/34/9/1207

Als, H., Duffy, F.H., McAnulty, G.B., 1988. Behavioral differences between preterm and full-term newborns as measured with the APIB system scores: I. Infant Behav. Dev. 11, 305–318. https://doi.org/10.1016/0163-6383(88)90016-1

Alshaikh, B., Yee, W., Lodha, A., Henderson, E., Yusuf, K., Sauve, R., 2014. Coagulase-negative staphylococcus sepsis in preterm infants and long-term neurodevelopmental outcome. J. Perinatol. Off. J. Calif. Perinat. Assoc. 34, 125–129. https://doi.org/10.1038/jp.2013.155

André, M., Lamblin, M.-D., d'Allest, A.M., Curzi-Dascalova, L., Moussalli-Salefranque, F., S Nguyen The, T., Vecchierini-Blineau, M.-F., Wallois, F., Walls-Esquivel, E., Plouin, P., 2010. Electroencephalography in premature and full-term infants. Developmental features and glossary. Neurophysiol. Clin. Clin. Neurophysiol. 40, 59–124. https://doi.org/10.1016/j.neucli.2010.02.002

Andriessen, P., Koolen, A.M.P., Berendsen, R.C.M., Wijn, P.F.F., ten Broeke, E.D.M., Oei, S.G., Blanco, C.E., 2003. Cardiovascular fluctuations and transfer function analysis in stable preterm infants. Pediatr. Res. 53, 89–97. https://doi.org/10.1203/00006450-200301000-00016

Ansari, A.H., De Wel, O., Lavanga, M., Caicedo, A., Dereymaeker, A., Jansen, K., Vervisch, J., De Vos, M., Naulaers, G., Van Huffel, S., 2018. Quiet sleep detection in preterm infants using deep convolutional neural networks. J. Neural Eng. 15, 066006. https://doi.org/10.1088/1741-2552/aadc1f

ANSeR- The Algorithm for Neonatal Seizure Recognition Study - Full Text View - ClinicalTrials.gov, URL https://clinicaltrials.gov/ct2/show/NCT02160171 (accessed 3.21.19).

Apgar, V., 1953. A proposal for a new method of evaluation of the newborn infant. Curr. Res. Anesth. Analg. 32, 260–267.

Aslan, K., Bozdemir, H., Şahin, C., Oğulata, S.N., Erol, R., 2008. A Radial Basis Function Neural Network Model for Classification of Epilepsy Using EEG Signals. J. Med. Syst. 32, 403–408. https://doi.org/10.1007/s10916-008-9145-9

Bai, S., Kolter, J.Z., Koltun, V., 2018. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. ArXiv180301271 Cs.

Ballot, D.E., Ramdin, T., Rakotsoane, D., Agaba, F., Davies, V.A., Chirwa, T., Cooper, P.A., 2017. Use of the Bayley Scales of Infant and Toddler Development, Third Edition, to Assess Developmental Outcome in Infants and Young Children in an Urban Setting in South Africa. Int. Sch. Res. Not. 2017. https://doi.org/10.1155/2017/1631760

Barrington, K.J., Stewart, S., Lee, 2002. Differing blood pressure thresholds in preterm infants, effects on frequency of diagnosis of hypotension and intraventricular haemorrhage. Pediatric Academic Societies Annual Meeting 2002, abstract. Baltimore, Maryland.

Bauer, K., Linderkamp, O., Versmold, H.T., 1993. Systolic blood pressure and blood volume in preterm infants. Arch. Dis. Child. 69, 521–522.

Bayley, N., 2006. Bayley Scales of Infant and Toddler Development, Third Edition: Administration Manual. ed. United States of America: Psychorp.

Becoming A Data-Driven CEO | Domo [WWW Document], URL https://www.domo.com/solution/data-never-sleeps-6 (accessed 3.6.19).

Bell, A.H., Greisen, G., Pryds, O., 1993. Comparison of the effects of phenobarbitone and morphine administration on EEG activity in preterm babies. Acta Paediatr. Oslo Nor. 1992 82, 35–39.

Bell, A.H., McClure, B.G., Hicks, E.M., 1990. Power spectral analysis of the EEG of term infants following birth asphyxia. Dev. Med. Child Neurol. 32, 990–998.

Bell, A.H., McClure, B.G., McCullagh, P.J., McClelland, R.J., 1991. Variation in power spectral analysis of the EEG with gestational age. J. Clin. Neurophysiol. Off. Publ. Am. Electroencephalogr. Soc. 8, 312–319.

Bengio, Y., Simard, P., Frasconi, P., 1994. Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Netw. 5, 157–166. https://doi.org/10.1109/72.279181

Biagioni, E., Frisone, M.F., Laroche, S., Kapetanakis, B.A., Ricci, D., Adeyi-Obe, M., Lewis, H., Kennea, N., Cioni, G., Cowan, F., Rutherford, M., Azzopardi, D., Mercuri, E., 2007. Maturation of cerebral electrical activity and development of cortical folding in young very preterm infants. Clin. Neurophysiol. Off. J. Int. Fed. Clin. Neurophysiol. 118, 53–59. https://doi.org/10.1016/j.clinph.2006.09.018

Bohanon, F.J., Mrazek, A.A., Shabana, M.T., Mims, S., Radhakrishnan, G.L., Kramer, G.C., Radhakrishnan, R.S., 2015. Heart Rate Variability Analysis is More Sensitive at Identifying Neonatal Sepsis than Conventional Vital Signs. Am. J. Surg. 210, 661–667. https://doi.org/10.1016/j.amjsurg.2015.06.002

Bollepalli, S.C., Challa, S.S., Jana, S., 2018. Robust Heartbeat Detection from Multimodal Data via CNN-based Generalizable Information Fusion. IEEE Trans. Biomed. Eng. https://doi.org/10.1109/TBME.2018.2854899

Botting, N., Powls, A., Cooke, R.W.I., Marlow, N., 1997. Attention Deficit Hyperactivity Disorders and Other Psychiatric Outcomes in Very Low Birthweight Children at 12 Years. J. Child Psychol. Psychiatry 38, 931–941. https://doi.org/10.1111/j.1469-7610.1997.tb01612.x

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., 2000. Randomizing Outputs to Increase Prediction Accuracy. Mach. Learn. 40, 229–242. https://doi.org/10.1023/A:1007682208299

Breiman, L., 1996. Bagging Predictors. Mach. Learn. 24, 123–140. https://doi.org/10.1023/A:1018054314350

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and regression trees, The Wadsworth statistics/probability series. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA.

Brown, C.J., Miller, S.P., Booth, B.G., Andrews, S., Chau, V., Poskitt, K.J., Hamarneh, G., 2014. Structural network analysis of brain development in young preterm neonates. NeuroImage 101, 667–680. https://doi.org/10.1016/j.neuroimage.2014.07.030

Cabal, L.A., Siassi, B., Zanini, B., Hodgman, J.E., Hon, E.E., 1980. Factors affecting heart rate variability in preterm infants. Pediatrics 65, 50–56.

Caicedo, A., Thewissen, L., Smits, A., Naulaers, G., Allegaert, K., Huffel, S.V., 2016. Changes in Oxygenation Levels Precede Changes in Amplitude of the EEG in Premature Infants, in: Oxygen Transport to Tissue XXXVIII, Advances in Experimental Medicine and Biology. Springer, Cham, 143–149. https://doi.org/10.1007/978-3-319-38810-6_19

Chai, R., Naik, G.R., Nguyen, T.N., Ling, S.H., Tran, Y., Craig, A., Nguyen, H.T., 2017. Driver Fatigue Classification With Independent Component by Entropy Rate Bound Minimization Analysis in an EEG-Based System. IEEE J. Biomed. Health Inform. 21, 715–724. https://doi.org/10.1109/JBHI.2016.2532354

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. ArXiv160302754 Cs 785–794. https://doi.org/10.1145/2939672.2939785

Chen, W., Wang, Y., Cao, G., Chen, G., Gu, Q., 2014. A random forest model based classification scheme for neonatal amplitude-integrated EEG. Biomed. Eng. OnLine 13, S4. https://doi.org/10.1186/1475-925X-13-S2-S4

Chen, W.-L., Kuo, C.-D., 2007. Characteristics of Heart Rate Variability Can Predict Impending Septic Shock in Emergency Department Patients with Sepsis. Acad. Emerg. Med. 14, 392–397. https://doi.org/10.1111/j.1553-2712.2007.tb01796.x

Cheng, W.-Y., Ou Yang, T.-H., Anastassiou, D., 2013a. Development of a prognostic model for breast cancer survival in an open challenge environment. Sci. Transl. Med. 5, 181ra50. https://doi.org/10.1126/scitranslmed.3005974

Cheng, W.-Y., Yang, T.-H.O., Anastassiou, D., 2013b. Biomolecular Events in Cancer Revealed by Attractor Metagenes. PLoS Comput. Biol. 9. https://doi.org/10.1371/journal.pcbi.1002920

Chollet, F., 2015. Keras. Deep Learning for humans. Keras.

Clarke, D.D., Sokoloff, L., 1999. Circulation and Energy Metabolism of the Brain. Basic Neurochem. Mol. Cell. Med. Asp. 6th Ed.

Coben, R., Mohammad-Rezazadeh, I., 2015. Neural Connectivity in Epilepsy as Measured by Granger Causality. Front. Hum. Neurosci. 9. https://doi.org/10.3389/fnhum.2015.00194

Cortes, C., Vapnik, V., 1995. Support-Vector Networks, in: Machine Learning. 273–297.

Costa, T., Boccignone, G., Ferraro, M., 2012. Gaussian Mixture Model of Heart Rate Variability. PLoS ONE 7. https://doi.org/10.1371/journal.pone.0037731

Cronin, C.M., Shapiro, C.R., Casiro, O.G., Cheang, M.S., 1995. The impact of very low-birth-weight infants on the family is long lasting. A matched control study. Arch. Pediatr. Adolesc. Med. 149, 151–158.

Dauphin, Y.N., Fan, A., Auli, M., Grangier, D., 2016. Language Modeling with Gated Convolutional Networks. ArXiv161208083 Cs.

Davis, A.S., Gantz, M.G., Do, B., Shankaran, S., Hamrick, S.E.G., Kennedy, K.A., Tyson, J.E., Chalak, L.F., Laptook, A.R., Goldstein, R.F., Hintz, S.R., Das, A., Higgins, R.D., Ball, M.B., Hale, E.C., Van Meurs, K.P., Eunice Kennedy Shriver National Institute of Child Health and Human Development Neonatal Research Network, 2015. Serial aEEG recordings in a cohort of extremely preterm infants: feasibility and safety. J. Perinatol. Off. J. Calif. Perinat. Assoc. 35, 373–378. https://doi.org/10.1038/jp.2014.217

DeGiorgio, C.M., Miller, P., Meymandi, S., Chin, A., Epps, J., Gordon, S., Gornbein, J., Harper, R.M., 2010. RMSSD, a Measure of Heart Rate Variability, Is Associated With Risk Factors For SUDEP: The SUDEP-7 Inventory. Epilepsy Behav. EB 19, 78–81. https://doi.org/10.1016/j.yebeh.2010.06.011

Dempsey, E.M., 2015. Under pressure to treat? Arch. Dis. Child. - Fetal Neonatal Ed. 100, F380–F381. https://doi.org/10.1136/archdischild-2015-308667

Dempsey, E.M., Barrington, K.J., 2009. Evaluation and Treatment of Hypotension in the Preterm Infant. Clin. Perinatol., Current Controversies in Perinatology 36, 75–85. https://doi.org/10.1016/j.clp.2008.09.003

Dempsey, E.M., Barrington, K.J., 2007a. Treating hypotension in the preterm infant: when and with what: a critical and systematic review. J. Perinatol. 27, 469–478. https://doi.org/10.1038/sj.jp.7211774

Dempsey, E.M., Barrington, K.J., 2007b. Treating hypotension in the preterm infant: when and with what: a critical and systematic review. J. Perinatol. 27, 469–478. https://doi.org/10.1038/sj.jp.7211774

Dempsey, E.M., Barrington, K.J., 2006. Diagnostic criteria and therapeutic interventions for the hypotensive very low birth weight infant. J. Perinatol. 26, 677–681. https://doi.org/10.1038/sj.jp.7211579

Dempsey, E.M., Barrington, K.J., Marlow, N., O'Donnell, C.P., Miletin, J., Naulaers, G., Cheung, P.-Y., Corcoran, D., Pons, G., Stranak, Z., Laere, D.V., Consortium, on behalf of the H., 2014. Management of Hypotension in Preterm Infants (The HIP Trial): A Randomised Controlled Trial of Hypotension Management in Extremely Low Gestational Age Newborns. Neonatology 105, 275–281. https://doi.org/10.1159/000357553

Dempsey, E.M., Hazzani, F.A., Barrington, K.J., 2009. Permissive hypotension in the extremely low birthweight infant with signs of good perfusion. Arch. Dis. Child. - Fetal Neonatal Ed. 94, F241–F244. https://doi.org/10.1136/adc.2007.124263

Deo, R.C., 2015. Machine Learning in Medicine. Circulation 132, 1920–1930. https://doi.org/10.1161/CIRCULATIONAHA.115.001593

Development of audit measures and guidelines for good practice in the management of neonatal respiratory distress syndrome. Report of a Joint Working Group of the British Association of Perinatal Medicine and the Research Unit of the Royal College of Physicians., 1992. . Arch. Dis. Child. 67, 1221–1227.

Dietterich, T.G., 2000. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. Mach. Learn. 40, 139–157. https://doi.org/10.1023/A:1007607513941

Dietterich, T.G., 1998. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Comput 10, 1895–1923. https://doi.org/10.1162/089976698300017197

Dimitrijević, L., Bjelaković, B., Čolović, H., Mikov, A., Živković, V., Kocić, M., Lukić, S., 2016. Assessment of general movements and heart rate variability in prediction of neurodevelopmental outcome in preterm infants. Early Hum. Dev. 99, 7–12. https://doi.org/10.1016/j.earlhumdev.2016.05.014

Doheny, K.K., Palmer, C., Browning, K.N., Jairath, P., Liao, D., He, F., Travagli, R.A., 2014. Diminished vagal tone is a predictive biomarker of necrotizing enterocolitis-risk in preterm infants. Neurogastroenterol. Motil. 26, 832–840. https://doi.org/10.1111/nmo.12337

Dorling, J.S., Field, D.J., Manktelow, B., 2005. Neonatal disease severity scoring systems. Arch. Dis. Child. - Fetal Neonatal Ed. 90, F11–F16. https://doi.org/10.1136/adc.2003.048488

Doyle, L.W., Faber, B., Callanan, C., Morley, R., 2003. Blood pressure in late adolescence and very low birth weight. Pediatrics 111, 252–257.

Draghici, A.E., Taylor, J.A., 2016. The physiological basis and measurement of heart rate variability in humans. J. Physiol. Anthropol. 35. https://doi.org/10.1186/s40101-016-0113-7

Draper, E.S., Manktelow, B., Field, D.J., James, D., 1999. Prediction of survival for preterm births by weight and gestational age: retrospective population based study. BMJ 319, 1093–1097.

Dumoulin, V., Visin, F., 2016. A guide to convolution arithmetic for deep learning. ArXiv160307285 Cs Stat.

Edmond, K., Zaidi, A., 2010. New Approaches to Preventing, Diagnosing, and Treating Neonatal Sepsis. PLoS Med. 7. https://doi.org/10.1371/journal.pmed.1000213

Electrophysiology, T.F. of the E.S. of C. the N.A.S. of P., 1996. Heart Rate Variability: Standards of Measurement, Physiological Interpretation, and Clinical Use. Circulation 93, 1043–1065. https://doi.org/10.1161/01.CIR.93.5.1043

Fairchild, K.D., Aschner, J.L., 2012. HeRO monitoring to reduce mortality in NICU patients [WWW Document]. Res. Rep. Neonatol. https://doi.org/10.2147/RRN.S32570

Far0n, 2018. xgbfi: XGBoost Feature Interactions & Importance.

Faust, O., Acharya, U.R., Tamura, T., 2012. Formal design methods for reliable computer-aided diagnosis: a review. IEEE Rev. Biomed. Eng. 5, 15–28. https://doi.org/10.1109/RBME.2012.2184750

Filippi, L., Pezzati, M., Poggi, C., Rossi, S., Cecchi, A., Santoro, C., 2007. Dopamine versus dobutamine in very low birthweight infants: endocrine effects. Arch. Dis. Child. Fetal Neonatal Ed. 92, F367–F371. https://doi.org/10.1136/adc.2006.098566

Finn, D., Boylan, G.B., Ryan, C.A., Dempsey, E.M., 2016. Enhanced Monitoring of the Preterm Infant during Stabilization in the Delivery Room. Front. Pediatr. 4. https://doi.org/10.3389/fped.2016.00030

Fleming, S., Thompson, M., Stevens, R., Heneghan, C., Plüddemann, A., Maconochie, I., Tarassenko, L., Mant, D., 2011. Normal ranges of heart rate and respiratory rate in children from birth to 18 years: a systematic review of observational studies. Lancet 377, 1011–1018. https://doi.org/10.1016/S0140-6736(10)62226-X

Freund, Y., Schapire, R.E., 1996. Experiments with a New Boosting Algorithm, in: Proceedings of the Thirteenth International Conference on International Conference on Machine Learning, ICML'96. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 148–156.

Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data Anal., Nonlinear Methods and Data Mining 38, 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Friedman, J.H., 1999. Stochastic Gradient Boosting. Comput. Stat. Data Anal. 38, 367–378.

Gagliardi, L., Cavazza, A., Brunelli, A., Battaglioli, M., Merazzi, D., Tandoi, F., Cella, D., Perotti, G.F., Pelti, M., Stucchi, I., Frisone, F., Avanzini, A., Bellù, R., 2004. Assessing mortality risk in very low birthweight infants: a comparison of CRIB, CRIB-II, and SNAPPE-II. Arch. Dis. Child. - Fetal Neonatal Ed. 89, F419–F422. https://doi.org/10.1136/adc.2003.031286

Godin, P.J., Buchman, T.G., 1996. Uncoupling of biological oscillators: a complementary hypothesis concerning the pathogenesis of multiple organ dysfunction syndrome. Crit. Care Med. 24, 1107–1116.

Goldberg, R.N., Chung, D., Goldman, S.L., Bancalari, E., 1980. The association of rapid volume expansion and intraventricular hemorrhage in the preterm infant. J. Pediatr. 96, 1060–1063. https://doi.org/10.1016/S0022-3476(80)80642-1

Golder, V., Hepponstall, M., Yiallourou, S.R., Odoi, A., Horne, R.S.C., 2013. Autonomic cardiovascular control in hypotensive critically ill preterm infants is impaired during the first days of life. Early Hum. Dev. 89, 419–423. https://doi.org/10.1016/j.earlhumdev.2012.12.010

Gómez-Herrero, G., Wu, W., Rutanen, K., Soriano, M., Pipa, G., Vicente, R., Gómez-Herrero, G., Wu, W., Rutanen, K., Soriano, M.C., Pipa, G., Vicente, R., 2015. Assessing Coupling Dynamics from an Ensemble of Time Series. Entropy 17, 1958–1970. https://doi.org/10.3390/e17041958

Goulding, R.M., Stevenson, N.J., Murray, D.M., Livingstone, V., Filan, P.M., Boylan, G.B., 2015. Heart rate variability in hypoxic ischemic encephalopathy: correlation with EEG grade and 2-y neurodevelopmental outcome. Pediatr. Res. 77, 681–687. https://doi.org/10.1038/pr.2015.28

Granger, C.W.J., 1969. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. Econometrica 37, 424–438. https://doi.org/10.2307/1912791

Greene, M.M., Patra, K., Nelson, M.N., Silvestri, J.M., 2012. Evaluating preterm infants with the Bayley-III: patterns and correlates of development. Res. Dev. Disabil. 33, 1948–1956. https://doi.org/10.1016/j.ridd.2012.05.024

Greenough, A., Cheeseman, P., Kavvadia, V., Dimitriou, G., Morton, M., 2002. Colloid infusion in the perinatal period and abnormal neurodevelopmental outcome in very low birth weight infants. Eur. J. Pediatr. 161, 319–323. https://doi.org/10.1007/s00431-002-0950-8

Greisen, G., 2005. Autoregulation of cerebral blood flow in newborn babies. Early Hum. Dev. 81, 423–428. https://doi.org/10.1016/j.earlhumdev.2005.03.005

Guan, M.Y., Gulshan, V., Dai, A.M., Hinton, G.E., 2017. Who Said What: Modeling Individual Labelers Improves Classification. ArXiv170308774 Cs.

Guyon, I., Elisseeff, A., 2003. An Introduction to Variable and Feature Selection. J. Mach. Learn. Res. 3, 1157–1182.

Hadase, M., Azuma, A., Zen, K., Asada, S., Kawasaki, T., Kamitani, T., Kawasaki, S., Sugihara, H., Matsubara, H., 2004. Very Low Frequency Power of Heart Rate Variability is a Powerful Predictor of Clinical Prognosis in Patients With Congestive Heart Failure. Circ. J. 68, 343–347. https://doi.org/10.1253/circj.68.343

Hallberg, B., Grossmann, K., Bartocci, M., Blennow, M., 2010. The prognostic value of early aEEG in asphyxiated infants undergoing systemic hypothermia treatment. Acta Paediatr. Oslo Nor. 1992 99, 531–536. https://doi.org/10.1111/j.1651-2227.2009.01653.x

Hanin, B., 2018. Which Neural Net Architectures Give Rise to Exploding and Vanishing Gradients?, in: Proceedings of the 32Nd International Conference on Neural Information Processing Systems, NIPS'18. Curran Associates Inc., USA, 580–589.

He, K., Zhang, X., Ren, S., Sun, J., 2015a. Deep Residual Learning for Image Recognition. ArXiv151203385 Cs.

He, K., Zhang, X., Ren, S., Sun, J., 2015b. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. ArXiv150201852 Cs.

Hicks, J.F., Fairchild, K.D., 2013. HeRO monitoring in the NICU : sepsis detection and beyond Heart rate observation (HeRO) monitoring was developed for detection of sepsis in preterm infants.

Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., Kingsbury, B., 2012. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. IEEE Signal Process. Mag. 29, 82–97. https://doi.org/10.1109/MSP.2012.2205597

Hjorth, B., 1970. EEG analysis based on time domain properties. Electroencephalogr. Clin. Neurophysiol. 29, 306–310. https://doi.org/10.1016/0013-4694(70)90143-4

Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20, 832–844. https://doi.org/10.1109/34.709601

Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory. Neural Comput. 9, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hoshen, Y., Weiss, R.J., Wilson, K.W., 2015. Speech acoustic modeling from raw multichannel waveforms, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4624–4628. https://doi.org/10.1109/ICASSP.2015.7178847

Hubert, L., Arabie, P., 1985. Comparing partitions. J. Classif. 2, 193–218. https://doi.org/10.1007/BF01908075

Ioffe, S., Szegedy, C., 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. ArXiv150203167 Cs.

Jayasinghe, D., Gill, A.B., Levene, M.I., 2003. CBF Reactivity in Hypotensive and Normotensive Preterm Infants. Pediatr. Res. 54, 848–853. https://doi.org/10.1203/01.PDR.0000088071.30873.DA

Jennekens, W., Niemarkt, H.J., Engels, M., Pasman, J.W., van Pul, C., Andriessen, P., 2012. Topography of maturational changes in EEG burst spectral power of the preterm infant with a normal follow-up at 2 years of age. Clin. Neurophysiol. Off. J. Int. Fed. Clin. Neurophysiol. 123, 2130–2138. https://doi.org/10.1016/j.clinph.2012.03.018

Jeong, J., Gore, J.C., Peterson, B.S., 2001. Mutual information analysis of the EEG in patients with Alzheimer's disease. Clin. Neurophysiol. Off. J. Int. Fed. Clin. Neurophysiol. 112, 827–835.

Jorch, G., Jorch, N., 1987. Failure of autoregulation of cerebral blood flow in neonates studied by pulsed Doppler ultrasound of the internal carotid artery. Eur. J. Pediatr. 146, 468–472.

Jost, K., Datta, A.N., Frey, U., Suki, B., Schulzke, S.M., 2016. Heart rate variability predicts duration of respiratory support in preterm infants. Eur. Respir. J. 48, PA1291. https://doi.org/10.1183/13993003.congress-2016.PA1291

Julien, C., 2006. The enigma of Mayer waves: Facts and models. Cardiovasc. Res. 70, 12–21. https://doi.org/10.1016/j.cardiores.2005.11.008

Kaczmarek, J., Chawla, S., Marchica, C., Dwaihy, M., Grundy, L., Sant'Anna, G.M., 2013. Heart rate variability and extubation readiness in extremely preterm infants. Neonatology 104, 42–48. https://doi.org/10.1159/000347101

Kandel, E.R., Schwartz, J.H., Thomas, J.M., 2000. Principles of Neural Science. New York: McGraw-Hill.

Katheria, A., Rich, W., Finer, N., 2012. Electrocardiogram provides a continuous heart rate faster than oximetry during neonatal resuscitation. Pediatrics 130, e1177-1181. https://doi.org/10.1542/peds.2012-0784

Kato, T., Okumura, A., Hayakawa, F., Tsuji, T., Natsume, J., Watanabe, K., 2011. Evaluation of brain maturation in pre-term infants using conventional and amplitude-integrated electroencephalograms. Clin. Neurophysiol. Off. J. Int. Fed. Clin. Neurophysiol. 122, 1967–1972. https://doi.org/10.1016/j.clinph.2010.12.063

Katura, T., Tanaka, N., Obata, A., Sato, H., Maki, A., 2006. Quantitative evaluation of interrelations between spontaneous low-frequency oscillations in cerebral hemodynamics and systemic cardiovascular dynamics. NeuroImage 31, 1592–1600. https://doi.org/10.1016/j.neuroimage.2006.02.010

Kauppila, M., Vanhatalo, S., Stevenson, N.J., 2018. Artifact detection in neonatal EEG using Gaussian mixture models, in: Eskola, H., Väisänen, O., Viik, J., Hyttinen, J. (Eds.), EMBEC & NBC 2017, IFMBE Proceedings. Springer Singapore, 221–224.

Kawahara, J., Brown, C.J., Miller, S.P., Booth, B.G., Chau, V., Grunau, R.E., Zwicker, J.G., Hamarneh, G., 2017. BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment. NeuroImage 146, 1038–1049. https://doi.org/10.1016/j.neuroimage.2016.09.046

Kinugasa, H., Hirayanagi, K., 1999. Effects of skin surface cooling and heating on autonomic nervous activity and baroreflex sensitivity in humans. Exp. Physiol. 84, 369–377.

Kistner, A., Celsi, G., Vanpee, M., Jacobson, S.H., 2000. Increased blood pressure but normal renal function in adult women born preterm. Pediatr. Nephrol. Berl. Ger. 15, 215–220.

Kluckow, M., Evans, N., 1996. Relationship between blood pressure and cardiac output in preterm infants requiring mechanical ventilation. J. Pediatr. 129, 506–512.

Korshunova, I., Kindermans, P., Degrave, J., Verhoeven, T., Brinkmann, B.H., Dambre, J., 2018. Towards Improved Design and Evaluation of Epileptic Seizure Predictors. IEEE Trans. Biomed. Eng. 65, 502–510. https://doi.org/10.1109/TBME.2017.2700086

Kraskov, A., Stögbauer, H., Grassberger, P., 2004. Estimating mutual information. Phys. Rev. E 69, 066138. https://doi.org/10.1103/PhysRevE.69.066138

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet Classification with Deep Convolutional Neural Networks, in: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS'12. Curran Associates Inc., USA, 1097–1105.

Lake, D.E., Richman, J.S., Griffin, M.P., Moorman, J.R., 2002. Sample entropy analysis of neonatal heart rate variability. Am. J. Physiol.-Regul. Integr. Comp. Physiol. 283, R789–R797. https://doi.org/10.1152/ajpregu.00069.2002

Lakshmanan, A., Agni, M., Lieu, T., Fleegler, E., Kipke, M., Friedlich, P.S., McCormick, M.C., Belfort, M.B., 2017. The impact of preterm birth <37 weeks on parents and families: a cross-sectional study in the 2 years after discharge from the neonatal intensive care unit. Health Qual. Life Outcomes 15, 38. https://doi.org/10.1186/s12955-017-0602-3

Larroque, B., Ancel, P.-Y., Marret, S., Marchand, L., André, M., Arnaud, C., Pierrat, V., Rozé, J.-C., Messer, J., Thiriez, G., Burguet, A., Picaud, J.-C., Bréart, G., Kaminski, M., EPIPAGE Study group, 2008. Neurodevelopmental disabilities and special care of 5-year-old children born before 33 weeks of gestation (the EPIPAGE study): a longitudinal cohort study. Lancet Lond. Engl. 371, 813–820. https://doi.org/10.1016/S0140-6736(08)60380-3

Laughon, M., Bose, C., Allred, E., O'Shea, T.M., Van Marter, L.J., Bednarek, F., Leviton, A., 2007. Factors Associated With Treatment for Hypotension in Extremely Low Gestational Age Newborns During the First Postnatal Week. Pediatrics 119, 273–280. https://doi.org/10.1542/peds.2006-1138

Lee, H., Pham, P., Largman, Y., Ng, A.Y., 2009. Unsupervised feature learning for audio classification using convolutional deep belief networks, in: Bengio, Y., Schuurmans, D., Lafferty, J.D., Williams, C.K.I., Culotta, A. (Eds.), Advances in Neural Information Processing Systems 22. Curran Associates, Inc., 1096–1104.

Lee, J., Park, J., Kim, K.L., Nam, J., 2017. Sample-level Deep Convolutional Neural Networks for Music Auto-tagging Using Raw Waveforms.

Lewis, L.D., Ching, S., Weiner, V.S., Peterfreund, R.A., Eskandar, E.N., Cash, S.S., Brown, E.N., Purdon, P.L., 2013. Local cortical dynamics of burst suppression in the anaesthetized brain. Brain J. Neurol. 136, 2727–2737. https://doi.org/10.1093/brain/awt174

Li, X., Chen, X., Yan, Y., Wei, W., Wang, Z.J., 2014. Classification of EEG Signals Using a Multiple Kernel Learning Support Vector Machine. Sensors 14, 12784–12802. https://doi.org/10.3390/s140712784

Lindner, M., Vicente, R., Priesemann, V., Wibral, M., 2011. TRENTOOL: A Matlab open source toolbox to analyse information flow in time series data with transfer entropy. BMC Neurosci. 12, 119. https://doi.org/10.1186/1471-2202-12-119

Lizier, J.T., 2014. JIDT: An information-theoretic toolkit for studying the dynamics of complex systems. Front. Robot. AI 1. https://doi.org/10.3389/frobt.2014.00011

Lizier, J.T., Heinzle, J., Horstmann, A., Haynes, J.-D., Prokopenko, M., 2011. Multivariate information-theoretic measures reveal directed information structure and task relevant changes in fMRI connectivity. J. Comput. Neurosci. 30, 85–107. https://doi.org/10.1007/s10827-010-0271-2

Lizier, J.T., Prokopenko, M., Zomaya, A.Y., 2012. Local measures of information storage in complex distributed computation. Inf. Sci. 208, 39–54. https://doi.org/10.1016/j.ins.2012.04.016

Lloyd, R., 2019. Electroencephalography of Premature Infants. Serial EEG analysis of preterm infants for the prediction of neurodevelopmental outcome at two years of age. (PhD thesis). University College Cork, Ireland.

Lloyd, R.O., O'Toole, J.M., Livingstone, V., Hutch, W.D., Pavlidis, E., Cronin, A.-M., Dempsey, E.M., Filan, P.M., Boylan, G.B., 2016. Predicting 2-y outcome in preterm infants using early multimodal physiological monitoring. Pediatr. Res. 80, 382–388. https://doi.org/10.1038/pr.2016.92

Löfhede, J., Löfgren, N., Thordstein, M., Flisberg, A., Kjellmer, I., Lindecrantz, K., 2008. Classification of burst and suppression in the neonatal electroencephalogram. J. Neural Eng. 5, 402–410. https://doi.org/10.1088/1741-2560/5/4/005

Lou, H.C., Lassen, N.A., Friis-Hansen, B., 1979. Impaired autoregulation of cerebral blood flow in the distressed newborn infant. J. Pediatr. 94, 118–121.

Lowe, J.R., Nolen, T.L., Vohr, B., Adams-Chapman, I., Duncan, A.F., Watterberg, K., 2013. Effect of primary language on developmental testing in children born extremely preterm. Acta Paediatr. Oslo Nor. 1992 102, 896–900. https://doi.org/10.1111/apa.12310

Lowen, S.B., Teich, M.C., 1996. The periodogram and Allan variance reveal fractal exponents greater than unity in auditory-nerve spike trains. Acoust. Soc. Am. J. 99, 3585–3591. https://doi.org/10.1121/1.414979

Mac Namee, B., Cunningham, P., Byrne, S., Corrigan, O.I., 2002. The problem of bias in training data in regression problems in medical decision support. Artif. Intell. Med. 24, 51–70. https://doi.org/10.1016/S0933-3657(01)00092-6

Mallard, C., Wang, X., 2012. Infection-induced vulnerability of perinatal brain injury. Neurol. Res. Int. 2012, 102153. https://doi.org/10.1155/2012/102153

Mangia, S., Giove, F., Tkáč, I., Logothetis, N.K., Henry, P.-G., Olman, C.A., Maraviglia, B., Di Salle, F., Uğurbil, K., 2009. Metabolic and hemodynamic events following changes in neuronal activity: current hypotheses, theoretical predictions and in vivo NMR experimental findings. J. Cereb. Blood Flow Metab. Off. J. Int. Soc. Cereb. Blood Flow Metab. 29, 441–463. https://doi.org/10.1038/jcbfm.2008.134

Manley, B.J., Dawson, J.A., Kamlin, C.O.F., Donath, S.M., Morley, C.J., Davis, P.G., 2010. Clinical assessment of extremely premature infants in the delivery room is a poor predictor of survival. Pediatrics 125, e559-564. https://doi.org/10.1542/peds.2009-1307

Mathieson, S.R., Stevenson, N.J., Low, E., Marnane, W.P., Rennie, J.M., Temko, A., Lightbody, G., Boylan, G.B., 2016. Validation of an automated seizure detection algorithm for term neonates. Clin. Neurophysiol. 127, 156–168. https://doi.org/10.1016/j.clinph.2015.04.075

Matić, V., Cherian, P.J., Widjaja, D., Jansen, K., Naulaers, G., Van Huffel, S., De Vos, M., 2013. Heart rate variability in newborns with hypoxic brain injury. Adv. Exp. Med. Biol. 789, 43–48. https://doi.org/10.1007/978-1-4614-7411-1_7

Ment, L.R., Duncan, C.C., Ehrenkranz, R.A., Lange, R.C., Taylor, K.J., Kleinman, C.S., Scott, D.T., Sivo, J., Gettner, P., 1984. Intraventricular hemorrhage in the preterm neonate: timing and cerebral blood flow changes. J. Pediatr. 104, 419–425.

Miotto, R., Wang, F., Wang, S., Jiang, X., Dudley, J.T., 2018. Deep learning for healthcare: review, opportunities and challenges. Brief. Bioinform. 19, 1236–1246. https://doi.org/10.1093/bib/bbx044

Morrell, L.K., Morrell, F., 1966. Evoked potentials and reaction times: a study of intra-individual variability. Electroencephalogr. Clin. Neurophysiol. 20, 567–575.

Murphy, K., Stevenson, N.J., Goulding, R.M., Lloyd, R.O., Korotchikova, I., Livingstone, V., Boylan, G.B., 2015. Automated analysis of multi-channel EEG in preterm infants. Clin. Neurophysiol. 126, 1692–1702. https://doi.org/10.1016/j.clinph.2014.11.024

Na, S.H., Jin, S.-H., Kim, S.Y., Ham, B.-J., 2002. EEG in schizophrenic patients: mutual information analysis. Clin. Neurophysiol. Off. J. Int. Fed. Clin. Neurophysiol. 113, 1954–1960.

Netoff, T.I., Carroll, T.L., Pecora, L.M., Schiff, S.J., 2006. Detecting Coupling in the Presence of Noise and Nonlinearity, in: Schelter, B., Winterhalder, tthias, Timmer, J. (Eds.), Handbook of Time Series Analysis. Wiley-VCH Verlag GmbH & Co. KGaA, 265–282. https://doi.org/10.1002/9783527609970.ch11

Ng, A.Y., Jordan, M.I., 2002. On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes, in: Dietterich, T.G., Becker, S., Ghahramani, Z. (Eds.), Advances in Neural Information Processing Systems 14. MIT Press, 841–848.

Ng, H.-W., Nguyen, V.D., Vonikakis, V., Winkler, S., 2015. Deep Learning for Emotion Recognition on Small Datasets Using Transfer Learning, in: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ICMI '15. ACM, New York, NY, USA, 443–449. https://doi.org/10.1145/2818346.2830593

Niedermeyer, E., Silva, F.H.L. da, 2005. Electroencephalography: Basic Principles, Clinical Applications, and Related Fields. Lippincott Williams & Wilkins.

Niemarkt, H.J., Andriessen, P., Peters, C.H.L., Pasman, J.W., Zimmermann, L.J., Bambang Oetomo, S., 2010. Quantitative analysis of maturational changes in EEG background activity in very preterm infants with a normal neurodevelopment at 1 year of age. Early Hum. Dev. 86, 219–224. https://doi.org/10.1016/j.earlhumdev.2010.03.003

Niemarkt, H.J., Jennekens, W., Pasman, J.W., Katgert, T., van Pul, C., Gavilanes, A.W.D., Kramer, B.W., Zimmermann, L.J., Bambang Oetomo, S., Andriessen, P., 2011. Maturational Changes in Automated EEG Spectral Power Analysis in Preterm Infants. Pediatr. Res. 70, 529–534. https://doi.org/10.1203/PDR.0b013e31822d748b

Nosek, B.A., Lakens, D., 2014. Registered Reports. Soc. Psychol. 45, 137–141. https://doi.org/10.1027/1864-9335/a000192

Odd, D.E., Rasmussen, F., Gunnell, D., Lewis, G., Whitelaw, A., 2008. A cohort study of low Apgar scores and cognitive outcomes. Arch. Dis. Child. - Fetal Neonatal Ed. 93, F115–F120. https://doi.org/10.1136/adc.2007.123745

O'Donnell, C.P.F., Kamlin, C.O.F., Davis, P.G., Carlin, J.B., Morley, C.J., 2006. Interobserver variability of the 5-minute Apgar score. J. Pediatr. 149, 486–489. https://doi.org/10.1016/j.jpeds.2006.05.040

Okumura, A., Kubota, T., Tsuji, T., Kato, T., Hayakawa, F., Watanabe, K., 2006. Amplitude Spectral Analysis of Theta/Alpha/Beta Waves in Preterm Infants. Pediatr. Neurol. 34, 30–34. https://doi.org/10.1016/j.pediatrneurol.2005.06.005

Omboni, S., Parati, G., Frattola, A., Mutti, E., Di Rienzo, M., Castiglioni, P., Mancia, G., 1993. Spectral and sequence analysis of finger blood pressure variability. Comparison with analysis of intra-arterial recordings. Hypertens. Dallas Tex 1979 22, 26–33.

Oord, A. van den, Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K., 2016. WaveNet: A Generative Model for Raw Audio. ArXiv160903499 Cs.

Osborn, D., Evans, N., Kluckow, M., 2002. Randomized trial of dobutamine versus dopamine in preterm infants with low systemic blood flow. J. Pediatr. 140, 183–191. https://doi.org/10.1067/mpd.2002.120834

Osborn, D.A., Evans, N., Kluckow, M., 2004. Clinical detection of low upper body blood flow in very premature infants using blood pressure, capillary refill time, and central-peripheral temperature difference. Arch. Dis. Child. - Fetal Neonatal Ed. 89, F168–F173. https://doi.org/10.1136/adc.2002.023796

O'Shea, A., Lightbody, G., Boylan, G., Temko, A., 2017. Neonatal Seizure Detection using Convolutional Neural Networks. ArXiv170905849 Cs Stat.

O'Toole, J.M., Boylan, G.B., Lloyd, R.O., Goulding, R.M., Vanhatalo, S., Stevenson, N.J., 2017. Detecting bursts in the EEG of very and extremely premature infants using a multi-feature approach. Med. Eng. Phys. 45, 42–50. https://doi.org/10.1016/j.medengphy.2017.04.003

Pagani, M., Lucini, D., Rimoldi, O., Furlan, R., Piazza, S., Porta, A., Malliani, A., 1996. Low and high frequency components of blood pressure variability. Ann. N. Y. Acad. Sci. 783, 10–23.

Pan, J., Tompkins, W.J., 1985. A real-time QRS detection algorithm. IEEE Trans. Biomed. Eng. 32, 230–236. https://doi.org/10.1109/TBME.1985.325532

Parry, G., Tucker, J., Tarnow-Mordi, W., 2003. CRIB II: an update of the clinical risk index for babies score. The Lancet 361, 1789–1791. https://doi.org/10.1016/S0140-6736(03)13397-1

Patel, M., Lal, S.K.L., Kavanagh, D., Rossiter, P., 2011. Applying neural network analysis on heart rate variability data to assess driver fatigue. Expert Syst. Appl. 38, 7235–7242. https://doi.org/10.1016/j.eswa.2010.12.028

Pavlidis, E., Lloyd, R.O., Livingstone, V., O'Toole, J.M., Filan, P.M., Boylan, G., 2019. A standardised assessment scheme for conventional EEG in preterm infants. Clinical Neurophysiology 131. https://doi.org/10.1016/j.clinph.2019.09.028

Pavlidis, E., Lloyd, R.O., Mathieson, S., Boylan, G.B., 2017. A review of important electroencephalogram features for the assessment of brain maturation in premature infants. Acta Paediatr. 106, 1394–1408. https://doi.org/10.1111/apa.13956

Pereda, E., Quiroga, R.Q., Bhattacharya, J., 2005. Nonlinear multivariate analysis of neurophysiological signals. Prog. Neurobiol. 77, 1–37. https://doi.org/10.1016/j.pneurobio.2005.10.003

Périvier, M., Rozé, J.-C., Gascoin, G., Hanf, M., Branger, B., Rouger, V., Berlie, I., Montcho, Y., Péréon, Y., Flamant, C., Tich, S.N.T., 2016. Neonatal EEG and neurodevelopmental outcome in preterm infants born before 32 weeks. Arch. Dis. Child. - Fetal Neonatal Ed. 101, F253–F259. https://doi.org/10.1136/archdischild-2015-308664

Pfurtscheller, G., Daly, I., Bauernfeind, G., Müller-Putz, G.R., 2012. Coupling between Intrinsic Prefrontal HbO2 and Central EEG Beta Power Oscillations in the Resting Brain. PLoS ONE 7. https://doi.org/10.1371/journal.pone.0043640

Piccirillo, G., Ogawa, M., Song, J., Chong, V.J., Joung, B., Han, S., Magrì, D., Chen, L.S., Lin, S.-F., Chen, P.-S., 2009. Power spectral analysis of heart rate variability and autonomic nervous system activity measured directly in healthy dogs and dogs with tachycardia-induced heart failure. Heart Rhythm 6, 546–552. https://doi.org/10.1016/j.hrthm.2009.01.006

Pierrat, V., Marchand-Martin, L., Arnaud, C., Kaminski, M., Resche-Rigon, M., Lebeaux, C., Bodeau-Livinec, F., Morgan, A.S., Goffinet, F., Marret, S., Ancel, P.-Y., EPIPAGE-2 writing group, 2017. Neurodevelopmental outcome at 2 years for preterm children born at 22 to 34 weeks' gestation in France in 2011: EPIPAGE-2 cohort study. BMJ 358, j3448. https://doi.org/10.1136/bmj.j3448

Pikovsky, A., Rosenblum, M., Kurths, J., 2003. Synchronization: A Universal Concept in Nonlinear Sciences [WWW Document]. Camb. Univ. Press. URL http://www.cambridge.org/ie/academic/subjects/physics/nonlinear-science-and-fluid-dynamics/synchronization-universal-concept-nonlinear-sciences (accessed 8.3.17).

Pincus, S.M., 1991. Approximate entropy as a measure of system complexity. Proc. Natl. Acad. Sci. U. S. A. 88, 2297–2301.

Pincus, S.M., Gladstone, I.M., Ehrenkranz, R.A., 1991. A regularity statistic for medical data analysis. J. Clin. Monit. 7, 335–345. https://doi.org/10.1007/BF01619355

Pincus, S.M., Keefe, D.L., 1992. Quantification of hormone pulsatility via an approximate entropy algorithm. Am. J. Physiol. 262, E741-754. https://doi.org/10.1152/ajpendo.1992.262.5.E741

Pincus, S.M., Viscarello, R.R., 1992. Approximate entropy: a regularity measure for fetal heart rate analysis. Obstet. Gynecol. 79, 249–255.

Pompe, B., Runge, J., 2011. Momentary information transfer as a coupling measure of time series. Phys. Rev. E 83, 051122. https://doi.org/10.1103/PhysRevE.83.051122

Pressler, R.M., Boylan, G.B., Morton, M., Binnie, C.D., Rennie, J.M., 2001. Early serial EEG in hypoxic ischaemic encephalopathy. Clin. Neurophysiol. 112, 31–37. https://doi.org/10.1016/S1388-2457(00)00517-4

Ragwitz, M., Kantz, H., 2002. Markov models from data by simple nonlinear time series predictors in delay embedding spaces. Phys. Rev. E 65, 056201. https://doi.org/10.1103/PhysRevE.65.056201

Rakow, A., Katz-Salamon, M., Ericson, M., Edner, A., Vanpée, M., 2013. Decreased heart rate variability in children born with low birth weight. Pediatr. Res. 74, 339–343. https://doi.org/10.1038/pr.2013.97

Rassi, D., Mishin, A., Zhuravlev, Y.E., Matthes, J., 2005. Time domain correlation analysis of heart rate variability in preterm neonates. Early Hum. Dev. 81, 341–350. https://doi.org/10.1016/j.earlhumdev.2004.09.002

Rennie, J., Boylan, G., 2007. Treatment of neonatal seizures. Arch. Dis. Child. Fetal Neonatal Ed. 92, F148–F150. https://doi.org/10.1136/adc.2004.068551

Rennie, J.M., Chorley, G., Boylan, G.B., Pressler, R., Nguyen, Y., Hooper, R., 2004. Non-expert use of the cerebral function monitor for neonatal seizure detection. Arch. Dis. Child. Fetal Neonatal Ed. 89, F37-40.

Reynolds, D.A., 2009. Gaussian Mixture Models, in: Encyclopedia of Biometrics. https://doi.org/10.1007/978-0-387-73003-5_196

Richardson, D.K., Phibbs, C.S., Gray, J.E., McCormick, M.C., Workman-Daniels, K., Goldmann, D.A., 1993. Birth weight and illness severity: independent predictors of neonatal mortality. Pediatrics 91, 969–975.

Roche-Labarbe, N., Wallois, F., Ponchel, E., Kongolo, G., Grebe, R., 2007. Coupled oxygenation oscillation measured by NIRS and intermittent cerebral activation on EEG in premature infants. NeuroImage 36, 718–727. https://doi.org/10.1016/j.neuroimage.2007.04.002

Rokach, L., 2010. Ensemble-based classifiers. Artif. Intell. Rev. 33, 1–39. https://doi.org/10.1007/s10462-009-9124-7

Rokach, L., Maimon, O., 2005. Top-down induction of decision trees classifiers - a survey. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 35, 476–487. https://doi.org/10.1109/TSMCC.2004.843247

Roulston, M.S., 1997. Significance testing of information theoretic functionals. Phys. Nonlinear Phenom. 110, 62–66. https://doi.org/10.1016/S0167-2789(97)00117-6

Schapire, R.E., 2003. The Boosting Approach to Machine Learning: An Overview, in: Denison, D.D., Hansen, M.H., Holmes, C.C., Mallick, B., Yu, B. (Eds.), Nonlinear Estimation and Classification, Lecture Notes in Statistics. Springer New York, 149–171. https://doi.org/10.1007/978-0-387-21579-2_9

Scher, M.S., 1996. Normal electrographic-polysomnographic patterns in preterm and fullterm infants. Semin. Pediatr. Neurol. 3, 2–12.

Schirrmeister, R., Gemein, L., Eggensperger, K., Hutter, F., Ball, T., 2017. Deep learning with convolutional neural networks for decoding and visualization of EEG pathology, in: 2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB). Presented at the 2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), 1–7. https://doi.org/10.1109/SPMB.2017.8257015

Schlüter, J., Böck, S., 2014. Improved musical onset detection with Convolutional Neural Networks, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6979–6983. https://doi.org/10.1109/ICASSP.2014.6854953

Schreiber, T., 2000. Measuring Information Transfer. Phys. Rev. Lett. 85, 461–464. https://doi.org/10.1103/PhysRevLett.85.461

Schumacher, E.M., Larsson, P.G., Sinding-Larsen, C., Aronsen, R., Lindeman, R., Skjeldal, O.H., Stiris, T.A., 2013. Automated spectral EEG analyses of premature infants during the first three days of life correlated with developmental outcomes at 24 months. Neonatology 103, 205–212. https://doi.org/10.1159/000345923

Schwartz, P.J., Garson, A., Paul, T., Stramba-Badiale, M., Vetter, V.L., Wren, C., European Society of Cardiology, 2002. Guidelines for the interpretation of the neonatal electrocardiogram. A task force of the European Society of Cardiology. Eur. Heart J. 23, 1329–1344.

Schwartz, P.J., Stramba-Badiale, M., Segantini, A., Austoni, P., Bosi, G., Giorgetti, R., Grancini, F., Marni, E.D., Perticone, F., Rosti, D., Salice, P., 1998. Prolongation of the QT interval and the sudden infant death syndrome. N. Engl. J. Med. 338, 1709–1714. https://doi.org/10.1056/NEJM199806113382401

Selig, F.A., Tonolli, E.R., Silva, E.V.C.M. da, Godoy, M.F. de, 2011. Heart rate variability in preterm and term neonates. Arq. Bras. Cardiol. 96, 443–449.

Semenova, O., Lightbody, G., O'Toole, J.M., Boylan, G., Dempsey, E., Temko, A., 2018. Coupling between mean blood pressure and EEG in preterm neonates is associated with reduced illness severity scores. PLOS ONE 13, e0199587. https://doi.org/10.1371/journal.pone.0199587

Semenova, O., Lightbody, G., O'Toole, J.M., Boylan, G., Dempsey, E., Temko, A., 2017. Modelling interactions between blood pressure and brain activity in preterm neonates. Conf. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf. 2017, 3969–3972. https://doi.org/10.1109/EMBC.2017.8037725

Shaffer, F., Ginsberg, J.P., 2017. An Overview of Heart Rate Variability Metrics and Norms. Front. Public Health 5. https://doi.org/10.3389/fpubh.2017.00258

Shah, A.J., Lampert, R., Goldberg, J., Veledar, E., Bremner, J.D., Vaccarino, V., 2013. Posttraumatic Stress Disorder and Impaired Autonomic Modulation in Male Twins. Biol. Psychiatry, Extinction and the Treatment of Anxiety Disorders 73, 1103–1110. https://doi.org/10.1016/j.biopsych.2013.01.019

Shah, D., Paradisis, M., Bowen, J.R., 2013. Relationship between systemic blood flow, blood pressure, inotropes, and aEEG in the first 48 h of life in extremely preterm infants. Pediatr. Res. 74, 314–320. https://doi.org/10.1038/pr.2013.104

Shaw, J.C., 1981. An introduction to the coherence function and its use in EEG signal analysis. J. Med. Eng. Technol. 5, 279–288.

Shelhamer, E., Long, J., Darrell, T., 2017. Fully Convolutional Networks for Semantic Segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39, 640–651. https://doi.org/10.1109/TPAMI.2016.2572683

Siewert-Delle, A., Ljungman, S., 1998. The impact of birth weight and gestational age on blood pressure in adult life: a population-based study of 49-year-old men. Am. J. Hypertens. 11, 946–953.

Simayijiang, Z., Backman, S., Ulén, J., Wikström, S., Åstrom, K., 2013. Exploratory study of EEG burst characteristics in preterm infants. Conf. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf. 2013, 4295–4298. https://doi.org/10.1109/EMBC.2013.6610495

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. ArXiv14091556 Cs.

Sinclair, D.B., Campbell, M., Byrne, P., Prasertsom, W., Robertson, C.M., 1999. EEG and long-term outcome of term infants with neonatal hypoxic-ischemic encephalopathy. Clin. Neurophysiol. Off. J. Int. Fed. Clin. Neurophysiol. 110, 655–659.

Sood, B.G., McLaughlin, K., Cortez, J., 2015. Near-infrared spectroscopy: applications in neonates. Semin. Fetal. Neonatal Med. 20, 164–172. https://doi.org/10.1016/j.siny.2015.03.008

Spencer-Smith, M.M., Spittle, A.J., Lee, K.J., Doyle, L.W., Anderson, P.J., 2015. Bayley-III Cognitive and Language Scales in Preterm Children. Pediatrics 135, e1258–e1265. https://doi.org/10.1542/peds.2014-3039

Spittle, A., Orton, J., Anderson, P.J., Boyd, R., Doyle, L.W., 2015. Early developmental intervention programmes provided post hospital discharge to prevent motor and cognitive impairment in preterm infants. Cochrane Database Syst. Rev. CD005495. https://doi.org/10.1002/14651858.CD005495.pub4

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J. Mach. Learn. Res. 15, 1929–1958.

Steinbrink, J., Villringer, A., Kempf, F., Haux, D., Boden, S., Obrig, H., 2006. Illuminating the BOLD signal: combined fMRI-fNIRS studies. Magn. Reson. Imaging 24, 495–505. https://doi.org/10.1016/j.mri.2005.12.034

Stéphan-Blanchard, E., Chardon, K., Léké, A., Delanaud, S., Bach, V., Telliez, F., 2013. Heart Rate Variability in Sleeping Preterm Neonates Exposed to Cool and Warm Thermal Conditions. PLOS ONE 8, e68211. https://doi.org/10.1371/journal.pone.0068211

Steuer, R., Kurths, J., Daub, C.O., Weise, J., Selbig, J., 2002. The mutual information: detecting and evaluating dependencies between variables. Bioinforma. Oxf. Engl. 18 Suppl 2, S231-240.

Stevenson, N.J., Tapani, K., Lauronen, L., Vanhatalo, S., 2019. A dataset of neonatal EEG recordings with seizure annotations. Sci. Data 6, 190039. https://doi.org/10.1038/sdata.2019.39

Stiles, J., Jernigan, T.L., 2010. The Basics of Brain Development. Neuropsychol. Rev. 20, 327–348. https://doi.org/10.1007/s11065-010-9148-4

Stober, S., Sternin, A., Owen, A.M., Grahn, J.A., 2015. Deep Feature Learning for EEG Recordings. ArXiv151104306 Cs.

Stoll, B.J., Hansen, N., Fanaroff, A.A., Wright, L.L., Carlo, W.A., Ehrenkranz, R.A., Lemons, J.A., Donovan, E.F., Stark, A.R., Tyson, J.E., Oh, W., Bauer, C.R., Korones, S.B., Shankaran, S., Laptook, A.R., Stevenson, D.K., Papile, L.-A., Poole, W.K., 2002. Late-onset sepsis in very low birth weight neonates: the experience of the NICHD Neonatal Research Network. Pediatrics 110, 285–291.

Stranak, Z., Semberova, J., Barrington, K., O'Donnell, C., Marlow, N., Naulaers, G., Dempsey, E., HIP consortium, 2014. International survey on diagnosis and management of hypotension in extremely preterm babies. Eur. J. Pediatr. 173, 793–798. https://doi.org/10.1007/s00431-013-2251-9

Suppiej, A., Cainelli, E., Cappellari, A., Trevisanuto, D., Balao, L., Di Bono, M.G., Bisiacchi, P.S., 2017. Spectral analysis highlight developmental EEG changes in preterm infants without overt brain damage. Neurosci. Lett. 649, 112–115. https://doi.org/10.1016/j.neulet.2017.04.021

Synnes, A.R., Chien, L.-Y., Peliowski, A., Baboolal, R., Lee, S.K., 2001. Variations in intraventricular hemorrhage incidence rates among Canadian neonatal intensive care units. J. Pediatr. 138, 525–531. https://doi.org/10.1067/mpd.2001.111822

Tataranno, M.L., Alderliesten, T., Vries, L.S. de, Groenendaal, F., Toet, M.C., Lemmers, P.M.A., De, R.E.V. van, Bel, F. van, Benders, M.J.N.L., 2015. Early Oxygen-Utilization and Brain Activity in Preterm Infants. PLOS ONE 10, e0124623. https://doi.org/10.1371/journal.pone.0124623

Temko, A., Doyle, O., Murray, D., Lightbody, G., Boylan, G., Marnane, W., 2015. Multimodal predictor of neurodevelopmental outcome in newborns with hypoxic-ischaemic encephalopathy. Comput. Biol. Med. 63, 169–177. https://doi.org/10.1016/j.compbiomed.2015.05.017

Temko, A., Thomas, E., Marnane, W., Lightbody, G., Boylan, G., 2011. EEG-based neonatal seizure detection with Support Vector Machines. Clin. Neurophysiol. 122, 464–473. https://doi.org/10.1016/j.clinph.2010.06.034

Thakor, N.V., Webster, J.G., Tompkins, W.J., 1984. Estimation of QRS complex power spectra for design of a QRS filter. IEEE Trans. Biomed. Eng. 31, 702–706. https://doi.org/10.1109/TBME.1984.325393

Thomaidis, C., Varlamis, G., Karamperis, S., 1988. Comparative Study of the Electrocardiograms of Healthy Fullterm and Premature Newborns. Acta Pædiatrica 77, 653–657. https://doi.org/10.1111/j.1651-2227.1988.tb10725.x

Thomas, E.M., Temko, A., Lightbody, G., Marnane, W.P., Boylan, G.B., 2009. A Gaussian mixture model based statistical classification system for neonatal seizure detection, in: 2009 IEEE International Workshop on Machine Learning for Signal Processing. Presented at the 2009 IEEE International Workshop on Machine Learning for Signal Processing, 1–6. https://doi.org/10.1109/MLSP.2009.5306203

Tokariev, A., Palmu, K., Lano, A., Metsäranta, M., Vanhatalo, S., 2012. Phase synchrony in the early preterm EEG: Development of methods for estimating synchrony in both oscillations and events. NeuroImage 60, 1562–1573. https://doi.org/10.1016/j.neuroimage.2011.12.080

Töllner, U., Bechinger, D., Pohlandt, F., 1980. Radial nerve palsy in a premature infant following long-term measurement of blood pressure. J. Pediatr. 96, 921–922. https://doi.org/10.1016/S0022-3476(80)80582-8

Tolonen, M., Palva, J.M., Andersson, S., Vanhatalo, S., 2007. Development of the spontaneous activity transients and ongoing cortical activity in human preterm babies. Neuroscience 145, 997–1006. https://doi.org/10.1016/j.neuroscience.2006.12.070

Turcott, R.G., Teich, M.C., 1996. Fractal character of the electrocardiogram: Distinguishing heart-failure and normal patients. Ann. Biomed. Eng. 24, 269–293. https://doi.org/10.1007/BF02667355

Tüske, Z., Golik, P., Schlüter, R., Ney, H., 2014. Acoustic modeling with deep neural networks using raw time signal for LVCSR, in: INTERSPEECH.

Tyszczuk, L., Meek, J., Elwell, C., Wyatt, J.S., 1998. Cerebral blood flow is independent of mean arterial blood pressure in preterm infants undergoing intensive care. Pediatrics 102, 337–341.

Van Marter, L.J., Leviton, A., Allred, E.N., Pagano, M., Kuban, K.C., 1990. Hydration during the first days of life and the risk of bronchopulmonary dysplasia in low birth weight infants. J. Pediatr. 116, 942–949.

Van Marter, L.J., Pagano, M., Allred, E.N., Leviton, A., Kuban, K.C.K., 1992. Rate of bronchopulmonary dysplasia as a function of neonatal intensive care practices. J. Pediatr. 120, 938–946. https://doi.org/10.1016/S0022-3476(05)81968-7

Vanhatalo, S., Palva, J.M., Andersson, S., Rivera, C., Voipio, J., Kaila, K., 2005. Slow endogenous activity transients and developmental expression of K+–Cl− cotransporter 2 in the immature human cortex. Eur. J. Neurosci. 22, 2799–2804. https://doi.org/10.1111/j.1460-9568.2005.04459.x

Vapnik, V., 2006. Estimation of Dependences Based on Empirical Data, Information Science and Statistics. Springer-Verlag, New York.

Vapnik, V., 1999. An overview of statistical learning theory, in: IEEE Transactions on Neural Networks 10.5. 988–999.

Vapnik, V.N., c1982. Estimation of dependences based on empirical data /. Springer-Verlag, New York :

Vargo, L., Seri, I., 2011. New NANN Practice Guideline: the management of hypotension in the very-low-birth-weight infant. Adv. Neonatal Care Off. J. Natl. Assoc. Neonatal Nurses 11, 272–278. https://doi.org/10.1097/ANC.0b013e318229263c

Vecchierini, M.-F., André, M., d'Allest, A.M., 2007. Normal EEG of premature infants born between 24 and 30 weeks gestational age: Terminology, definitions and maturation aspects. Neurophysiol. Clin. Neurophysiol. 37, 311–323. https://doi.org/10.1016/j.neucli.2007.10.008

Vesoulis, Z.A., Hao, J., McPherson, C., El Ters, N.M., Mathur, A.M., 2017. Low-frequency blood pressure oscillations and inotrope treatment failure in premature infants. J. Appl. Physiol. Bethesda Md 1985 123, 55–61. https://doi.org/10.1152/japplphysiol.00205.2017

Vicente, R., Wibral, M., Lindner, M., Pipa, G., 2011. Transfer entropy—a model-free measure of effective connectivity for the neurosciences. J. Comput. Neurosci. 30, 45–67. https://doi.org/10.1007/s10827-010-0262-3

Victor, S., Appleton, R.E., Beirne, M., Marson, A.G., Weindling, A.M., 2006a. The Relationship between Cardiac Output, Cerebral Electrical Activity, Cerebral Fractional Oxygen Extraction and Peripheral Blood Flow in Premature Newborn

Infants. Pediatr. Res. 60, 456–460. https://doi.org/10.1203/01.pdr.0000238379.67720.19

Victor, S., Marson, A.G., Appleton, R.E., Beirne, M., Weindling, A.M., 2006b. Relationship Between Blood Pressure, Cerebral Electrical Activity, Cerebral Fractional Oxygen Extraction, and Peripheral Blood Flow in Very Low Birth Weight Newborn Infants. Pediatr. Res. 59, 314–319. https://doi.org/10.1203/01.pdr.0000199525.08615.1f

Vinh, N.X., Epps, J., Bailey, J., 2010. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. J Mach Learn Res 11, 2837–2854.

Vinh, N.X., Epps, J., Bailey, J., 2009. Information Theoretic Measures for Clusterings Comparison: Is a Correction for Chance Necessary?, in: Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09. ACM, New York, NY, USA, 1073–1080. https://doi.org/10.1145/1553374.1553511

Vohr, B., 2014. Speech and language outcomes of very preterm infants. Semin. Fetal. Neonatal Med., Long-term outcome for the tiniest or most immature babies 19, 78–83. https://doi.org/10.1016/j.siny.2013.10.007

Vollmer, B., Roth, S., Baudin, J., Stewart, A.L., Neville, B.G.R., Wyatt, J.S., 2003. Predictors of Long-Term Outcome in Very Preterm Infants: Gestational Age Versus Neonatal Cranial Ultrasound. Pediatrics 112, 1108–1114. https://doi.org/10.1542/peds.112.5.1108

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., Maas, H.L.J. van der, Kievit, R.A., 2012. An Agenda for Purely Confirmatory Research. Perspect. Psychol. Sci. 7, 632–638. https://doi.org/10.1177/1745691612463078

Walch, E., Chaudhary, T., Herold, B., Obladen, M., 2009. Parental bilingualism is associated with slower cognitive development in very low birth weight infants. Early Hum. Dev. 85, 449–454. https://doi.org/10.1016/j.earlhumdev.2009.03.002

Wallois, F., 2010. Synopsis of maturation of specific features in EEG of premature neonates. Neurophysiol. Clin. Neurophysiol. 40, 125–126. https://doi.org/10.1016/j.neucli.2010.02.001

Weindling, A.M., 1989. Blood pressure monitoring in the newborn. Arch. Dis. Child. 64, 444–447.

Welch, P., 1967. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. IEEE Trans. Audio Electroacoustics 15, 70–73. https://doi.org/10.1109/TAU.1967.1161901

Werbos, P.J., 1975. Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. Harvard University.

West, C.R., Groves, A.M., Williams, C.E., Harding, J.E., Skinner, J.R., Kuschel, C.A., Battin, M.R., 2006. Early Low Cardiac Output Is Associated with Compromised Electroencephalographic Activity in Very Preterm Infants. Pediatr. Res. 59, 610–615. https://doi.org/10.1203/01.pdr.0000203095.06442.ad

White, D.M., Cott, C.A.V., 2010. EEG Artifacts in the Intensive Care Unit Setting. Am. J. Electroneurodiagnostic Technol. 50, 8–25. https://doi.org/10.1080/1086508X.2010.11079750

WHO | Preterm birth [WWW Document], n.d. . WHO. URL http://www.who.int/mediacentre/factsheets/fs363/en/ (accessed 8.23.16).

Wibral, M., Pampu, N., Priesemann, V., Siebenhühner, F., Seiwert, H., Lindner, M., Lizier, J.T., Vicente, R., 2013. Measuring Information-Transfer Delays. PLOS ONE 8, e55809. https://doi.org/10.1371/journal.pone.0055809

Wiesler, S., Ney, H., 2011. A Convergence Analysis of Log-Linear Training, in: Shawe-Taylor, J., Zemel, R.S., Bartlett, P.L., Pereira, F., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 24. Curran Associates, Inc., 657–665.

Wikström, S., Pupp, I.H., Rosén, I., Norman, E., Fellman, V., Ley, D., Hellström-Westas, L., 2012. Early single-channel aEEG/EEG predicts outcome in very preterm infants. Acta Paediatr. Oslo Nor. 1992 101, 719–726. https://doi.org/10.1111/j.1651-2227.2012.02677.x

Wollstadt, P., Sellers, K.K., Rudelt, L., Priesemann, V., Hutt, A., Fröhlich, F., Wibral, M., 2017. Breakdown of local information processing may underlie isoflurane anesthesia effects. PLOS Comput. Biol. 13, e1005511. https://doi.org/10.1371/journal.pcbi.1005511

Wulsin, D.F., Gupta, J.R., Mani, R., Blanco, J.A., Litt, B., 2011. Modeling electroencephalography waveforms with semi-supervised deep belief nets: fast classification and anomaly measurement. J. Neural Eng. 8, 036015. https://doi.org/10.1088/1741-2560/8/3/036015

Yamada, T., Meng, E., 2012. Practical Guide for Clinical Neurophysiologic Testing: EEG. Lippincott Williams & Wilkins.

Yu, F., Koltun, V., 2015. Multi-Scale Context Aggregation by Dilated Convolutions. ArXiv151107122 Cs.

Zhang, J., Penny, D.J., Kim, N.S., Yu, V.Y.H., Smolich, J.J., 1999. Mechanisms of blood pressure increase induced by dopamine in hypotensive preterm neonates. Arch. Dis. Child. - Fetal Neonatal Ed. 81, F99–F104. https://doi.org/10.1136/fn.81.2.F99

Zhang, J., Yan, C., Gong, X., 2017. Deep convolutional neural network for decoding motor imagery based brain computer interface, in: 2017 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC). Presented at the 2017 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), 1–5. https://doi.org/10.1109/ICSPCC.2017.8242581

Zweig, M.H., Campbell, G., 1993. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin. Chem. 39, 561–577.